



IEEE

TRANSACTIONS ON

AUTOMATIC CONTROL

A PUBLICATION OF THE IEEE CONTROL SYSTEMS SOCIETY

JANUARY 1995

VOLUME 40

NUMBER 1

IETAA9

(ISSN 0018-9286)

Scanning the Issue

1

PAPERS

Percentile Performance Criteria For Limiting Average Markov Decision Processes	<i>J. A. Eilar, D. Krass, and K. W. Ross</i>	2
Design and Analysis of Fuzzy Identifiers of Nonlinear Dynamic Systems	<i>J. A. Wang</i>	11
Stochastic Approximation with Averaging and Feedback: Rapidly Convergent On-Line Algorithms	<i>H. T. Kushner and J. Yang</i>	24
Exponential Stabilization of Nonholonomic Chained Systems	<i>O. J. Sordalen and O. Egeland</i>	35
Global Total Least Squares Modeling of Multivariable Time Series	<i>B. Roorda and C. Heij</i>	50
Control of Chained Systems: Application to Path Following and Time Varying Point Stabilization of Mobile Robots	<i>C. Samson</i>	64

TECHNICAL NOTES AND CORRESPONDENCE

Input-Output Robust Tracking Control Design for Flexible Joint Robots	<i>Z. Qu</i>	78
The Use of Symbolic Computation in Nonlinear Control: Is It Viable?	<i>B. de laet</i>	84
Note on Decentralized Adaptive Controller Design	<i>J. Lyou</i>	89
Characterization of Zeros in Two-Frequency Scale Systems	<i>H. M. Oloomi and M. F. Sawan</i>	92
Semi-Global Stabilizability of Linear Null Controllable Systems with Input Nonlinearities	<i>A. R. Tecl</i>	96
Stability and Exponential Stability of an Adaptive Control Scheme for Plants of any Relative Degree	<i>G. Bartolini, A. Ferrara, and A. A. Stotsky</i>	100
Preservation of Reachability and Observability under Sampling with a First Order Hold	<i>T. Hagiwara</i>	104
On Robust Asymptotic Tracking: Perturbations on Coprime Factors and Parameterization of All Solutions	<i>G. O. Cornea and M. A. da Silva</i>	107
A Version of Hautus' Test for Tandem Connection of Linear Systems	<i>A. Bacciotti and G. Beccati</i>	111
Poisson LQR Design for Asynchronous Multirate Controllers	<i>R. P. Tecland</i>	115
On Solving Diophantine Equations by Real Matrix Manipulation	<i>M. Yamada, P. C. Zuo, and Y. Funahashi</i>	118
Nonlinear Versus Linear Control in the Absolute Stabilizability of Uncertain Systems with Structured Uncertainty	<i>A. A. Savkin and I. R. Petersen</i>	122

(Continued on back cover)



The Control Systems Society is an organization within the framework of the IEEE of members with professional interest in automatic control. All members of the IEEE are eligible for membership in the Society and will receive the TRANSACTIONS upon payment of the annual Society membership fee of \$15.00. For further information write to the address below. Member copies of Transactions/Journals are for personal use only.

IEEE TRANSACTIONS ON AUTOMATIC CONTROL

Founding Editor: George S. Axelby

EDITORIAL BOARD

J. BAILEY *Editor in Chief*
Dep. Aerospace/Mechanical Eng.
Boston Univ.
110 Cummington St.
Boston, MA 02215
(617) 353-9848

C. CASSANDRAS *Editor Tech. Notes and Corresp.*
Dep. Electrical/Computer Eng.
Univ. Massachusetts
Amherst, MA 01003
(413) 545-0870

D. S. NAIK *Assoc. Editor Book Reviews*
College Eng.
Idaho State Univ.
871 South Eighth St.
Pocatello, ID 83209
(208) 236-2307

Associate Editors, Technical

M. ABADI
Kyoto Univ.
A. BAGCHI
Twente Univ. Technol.
A. M. BLOCH
Ohio State Univ.
J. A. BURNS
VPI & SU
C. CANUDAS DE WIT
ENSIEG INPG
E. K. P. CHONG
Purdue Univ.
M. DAHLHJ
Univ. Calif. Santa Barbara
M. DAHLHJ
Mass. Inst. Technol.
W. P. DAYAWANNA
Univ. Maryland

M. FU
Univ. Newcastle
W. B. GONG
Univ. Massachusetts
N. GUNDES
Univ. Calif. Davis
J. HALPERN
Univ. Colorado
C. V. HOLLIST
Univ. Massachusetts
G. HIRSH
Texas A&M Univ.
I. JANBARI
Univ. Calif. Irvine
S. LAHORITZ
Univ. Michigan
J. LASPERRI
CNRS

B. LITMAN
Mississippi State
P. NAIN
INRIA
R. NIKOLKHAH
INRIA
K. PASSINO
Ohio State Univ.
J. B. POMEI
INRIA
T. A. POSBERG
Univ. Minnesota
W. RIN
Univ. Calif. Berkeley
A. SAMRI
Washington State Univ.
J. SHAMMA
Univ. Texas, Austin

V. SOTTO
Macquarie Univ.
A. A. STOKROVITZ
Eindhoven Univ. Technol.
J. SUN
Ford Research Lab.
A. TESI
Univ. di Firenze
A. J. VAN DER SCHALF
Univ. Twente
P. VAN DOORTEN
Louvain Univ.
A. VICINO
Universita di Siena
E. YAZ
Univ. Arkansas
C. YIN
Wayne State Univ.

Associate Editors at Large

A. BENVENISTE
IRISA/INRIA
P. CROUCH
Arizona State Univ.
D. L. ELIOTT
Univ. Maryland

A. ISHORI
Univ. Roma La Sapienza &
Washington Univ.
P. R. KILMAR
Univ. Illinois Urbana
I. LJUNG
Linköping Univ.

A. N. MICHEL
Univ. Notre Dame
S. K. MITTER
Mass. Inst. Technol.
M. MORARI
Calif. Inst. Technol.

M. K. SAIN
Univ. Notre Dame
F. D. SONTAG
Rutgers Univ.
A. TITS
Univ. of Maryland

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

Officers

JAMES T. CAIN *President*
WALLACE S. READ *President Elect*
CHARLES W. TURNER *Secretary*
V. THOMAS RIVINS *Treasurer*
KENNETH R. LAKER *Vice President Educational Activities*

TSYU JONG (I. J.) TARN *Dist. for Division X—Systems and Control Division*

JOEL B. SNYDER *Vice President Professional Activities*
W. KENNETH DAWSON *Vice President Publication Activities*
VIJAY K. BHARADWAJ *Vice President Regional Activities*
F. G. KILNER *Vice President Standards Activities*
BRUCE EISENSTEIN *Vice President Technical Activities*

Executive Staff

THEODORE W. HINSEY *Executive Director*
RICHARD D. SCHWARTZ *Acting General Manager*
PHYLLIS HALL *Staff Executive Publications*
FRANK R. MOORE *Staff Executive Volunteer Activities*

IRVING ENGELSON *Staff Director Corporate Activities*
W. R. HABINGREITER *Staff Director Customer Service Center*
PETER A. LEWIS *Staff Director Educational Activities*
MEVIN L. OIKEN *Staff Director Regional Activities*

ANDREW G. SALIM *Staff Director Standards Activities*
W. THOMAS SUTLI *Staff Director Professional Activities*
ROBERT T. WANGELMANN *Staff Director Technical Activities*

Transactions/Journals Department

Director PATRICIA WALKER
Manager GAIL S. TERRELL
Electronic Production Manager JEFF UZZO
Managing Editor VALERIE CAMMARATA
Senior Associate Editor GERARDINE F. KROHN *Associate Editor* DAWN SPETH WHITE

IEEE TRANSACTIONS ON AUTOMATIC CONTROL is published monthly by The Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon IEEE, the Society, Council of its members. **IEEE Corporate Office** 345 East 47th Street, New York, NY 10017-2394. **IEEE Operations Center** 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. **NJ Telephone** 908-981-0060. **Price/Publication Information** Individual copies: IEEE Members \$10.00 (first copy only); nonmembers \$20.00 per copy. (Note: Add \$4.00 postage and handling charge to any order from \$1.00 to \$50.00 including prepaid orders.) Member and nonmember subscription prices available on request. Available in microfilm and microfiche. **Copyright and Reprint Permissions** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons provided the per copy fee indicated at the bottom of the first page is paid through the Copyright Clearance Center, 29 Congress Street, Salem, MA 01970. For all other copying, reprint, or republication permission write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. Copyright ©1995 by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Second class postage paid at New York, NY and at additional mailing offices. **Postmaster** Send address changes to IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

(Continued from front cover)

A Linear Algebraic Framework for Dynamic Feedback Linearization . . .	<i>E. Aranda-Bricaire, C. H. Moog, and J.-B. Pomet</i>	127
Design of a Class of Luenberger Observers for Descriptor Systems . . .	<i>M. Hou and P. C. Müller</i>	133
Comments on "On the Robust Popov Criterion for Interval Lur'e Systems" . . .	<i>T. Mori, T. Nishimura, Y. Kuroe, and H. Kokame</i>	136
An Output Feedback Globally Stable Controller for Induction Motors . . .	<i>G. Espinosa-Pérez and R. Ortega</i>	138
Eigenstructure Assignment in Linear Descriptor Systems . . .	<i>P. Zagulak and V. Kučera</i>	144
Robust Regulation in the Presence of Norm-Bounded Uncertainty . . .	<i>J. Abedor, K. Nagpal, P. P. Khargonekar, and K. Poolla</i>	147
Frequency-Domain Criteria of Robust Stability for Slowly Time-Varying Systems . . .	<i>A. Megretski</i>	153
Routing and Scheduling in Heterogeneous Systems: A Sample Path Approach . . .	<i>P. D. Sparaggis</i>	156
Methods and Theory for Off-Line Machine Learning . . .	<i>S. Yakowitz and J. Mai</i>	161
Discrete-Time Observers with Random Noises in Dynamic Block . . .	<i>E. A. Lyashenko and L. B. Ryashko</i>	165
Nevanlinna–Pick Interpolation Problem for Two Frequency Scale Systems . . .	<i>H. M. Oloomi</i>	169
Multiscale Smoothing Error Models . . .	<i>M. R. Luetgen and A. S. Willsky</i>	173
A Generalized Popov–Belevitch–Hautus Test of Observability . . .	<i>B. K. Ghosh and J. Rosenthal</i>	176
A Note on Robust Pole Placement . . .	<i>M. K. Solak and A. C. Peng</i>	181
Pole Assignment for Uncertain Systems in a Specified Disk by State Feedback . . .	<i>G. Garcia and J. Bernussou</i>	184
Modified Output Error Identification—Elimination of the SPR Condition . . .	<i>A. Betser and E. Zehch</i>	190
Comments on "Explicit Asymmetric Bounds for Robust Stability of Continuous and Discrete-Time Systems" . . .	<i>J. Xiao</i>	194
IEEE Copyright Information . . .		195

IEEE CONTROL SYSTEMS SOCIETY

BOARD OF GOVERNORS

Executive Officers

D. P. ATHERTON, President
Univ. Sussex
Brighton, UK BN1 9QT

M. K. MASTEN, President Elect
Texas Instruments
Plano, TX 75075

J. D. BIRDWELL, Secretary-Administrator
Univ. Tennessee
Knoxville, TN 37996

P. J. ANTSAKLIS, Vice President Conference Activities
Univ. Notre Dame
Notre Dame, IN 46556

N. H. McCLAMROCH, Vice President Financial Affairs
Univ. Michigan
Ann Arbor, MI 48109

D. P. LORZI, Vice President Member Activities
Univ. Massachusetts
Amherst, MA 01003

A. N. MICHEL, Vice President Technical Affairs
Univ. Notre Dame
Notre Dame, IN 46556

S. YURKOVICH, Vice President Publication Activities
Ohio State Univ.
Columbus, OH 43210

J. BAILEY, Editor in Chief
Transactions on Automatic Control

B. H. KROGH, Editor
Transactions on Control Systems Technology

S. YURKOVICH, Editor
IEEE Control Systems Magazine

Members

Term Ending December 31, 1995

*Term Ending December 31, 1995
(Appointed)*

Term Ending December 31, 1996

Term Ending December 31, 1997

J. L. ACKERMANN
P. E. CAINE
P. K. HOUP
N. H. McCLAMROCH
A. N. MICHEL
M. P. POLIS

G. DUMONT
C. SCHRAEDER
M. SPONG
B. SRIDHAR
S. VITTAL RAO
G. GOODWIN

P. J. ANTSAKLIS
K. FURUTA
T. L. JOHNSON
M. K. MASTEN
I. SHAW
S. YURKOVICH

A. ANNASWAMY
H. KIMURA
B. KROGH
K. PASSINO
D. REPPINGER
M. GIEVERS

INFORMATION FOR AUTHORS

In the IEEE TRANSACTIONS ON AUTOMATIC CONTROL the IEEE Control Systems Society publishes high-quality papers on the theory, design, and application of control systems. The TRANSACTIONS is published monthly.

Two types of contributions are regularly considered:

- 1) *Papers*—Presentation of significant research, development, or application of control concepts.
- 2) *Technical Notes and Correspondence*—Brief technical notes, comments on published areas or established control topics, corrections to papers and notes published in the TRANSACTIONS.

In addition, special papers (tutorials, surveys, and perspectives on the theory and applications of control systems topics) are solicited. Authors are urged to contact one of the Associate Editors at Large before submitting such papers.

Submission of a manuscript signifies that it has been neither copyrighted, published, nor submitted or accepted for publication elsewhere.

Submitted manuscripts must be typewritten in English. All submitted manuscripts should be as concise as possible. Technical Notes and Correspondence are normally limited to 12 double-spaced, typed pages. Papers of length exceeding 40 double-spaced, typed pages are strongly discouraged. The Editor reserves the right to refuse to consider such papers.

A. Process for Submission of Manuscript

- 1) *Papers*—Send seven copies of the paper together with two copies of a transmittal letter enclosed in one package to the Editor in Chief. These will be subject to a full review procedure, and a decision on whether or not to accept the paper will be made by members of the Transactions Editorial Board.
- 2) *Technical Notes and Correspondence*—Technical Notes and Correspondence should be sent to the Editor for Technical Notes and Correspondence; five copies are required. Decisions are made on the basis of a simplified review procedure.
- 3) Original illustrations should not be sent until requested, but authors should be ready to submit these immediately upon acceptance for publication.
- 4) Authors of accepted Papers are requested to supply their biographies (100 words or less) and photographs. For style, see biographies in this issue.
- 5) Enclose two copies of your letter of transmittal. Do not send letters of transmittal under separate cover. Give your preferred address for correspondence. Inform the Editor in Chief or Editor for Technical Notes and Correspondence of any change of address.
- 6) Authors of accepted manuscripts will be required to provide the final version of their manuscript on a computer diskette along with the hard copy.

B. Copyright

It is the policy of the IEEE to own the copyright to the technical contribution it publishes on behalf of the interests of the IEEE, its authors, and their employers, and to facilitate the appropriate reuse of this material by others. To comply with the U.S. Copyright Law, authors are required to sign an IEEE copyright form before publication. This form, a copy of which appears in the January 1995 issue of this journal, returns to authors and their employers full rights to reuse their material for their own purposes. Authors must submit a signed copy of this form with their manuscripts.

C. Style for Manuscript

- 1) First page must contain: a) Title of paper (without symbols); b) Author(s) and affiliation(s); c) Abstract (not exceeding 300 words for Papers or 75 words for Technical Notes and Correspondence, and without equations, references, or footnotes); d) Complete mailing address, telephone number, and if available, electronic mail (email) address and facsimile (fax) number of each author; e) Preferred address for correspondence and return of proofs; and f) Footnotes (if desired) containing acknowledgment of financial or other support.
- 2) If possible, please prepare copies using two sides of each page. Authors should be aware, however, that after acceptance of a manuscript, a one-sided copy must be provided for IEEE production purposes.
- 3) Provide an Introduction that includes a statement of the purpose and contribution of the paper.
- 4) If appropriate, indicate advantages, limitations, and possible applications in a Conclusion section.
- 5) References should be numbered and appear in a separate bibliography at the end of the paper. Use numerals in square brackets to cite references, e.g., [15]. References should be complete and in IEEE style (see examples in this issue).

D. Style for Illustrations

- 1) It is in the author's interest to submit professional quality illustrations. Drafting or art service cannot be provided by the IEEE.
- 2) Original drawings should be in black ink on white background. Maximum size is restricted to 21.6 by 27.9 cm. Glossy prints of illustrations are also acceptable.
- 3) All lettering should be large enough to permit legible reduction of the figure to column width, sometimes as small as one quarter of the original size. Typed lettering is usually not acceptable on figures.
- 4) Lightly pencil each figure number on the back of each original illustration. Captions should not appear on figures.
- 5) Provide a separate sheet listing all figure captions in proper style for the typesetter, e.g., Fig. 5. The error variance for the optimal filter.
- 6) Contributors' photographs should measure between 1.6 cm and 9.5 cm across the widest part of the head. The overall size of the photographic paper used can be anything from passport size to 27.9 cm.

E. Mandatory Overlength Page Charges

A mandatory page charge is imposed on all papers exceeding eight TRANSACTIONS pages (about 28 double-spaced, typed pages) in length, including illustrations. The charge is \$125 per page for each page over the first eight and is a prerequisite for publication. Details are provided at the time of acceptance; authors are, however, urged to keep this in mind when submitting and revising their papers.

F. Page Charges

After a manuscript has been accepted for publication, the author's company or institution will be approached with a request to pay a charge of \$110 per page to cover part of the cost of publication. Payment of page charges for this IEEE TRANSACTIONS, like journals of other professional societies, is not a necessary prerequisite for publication. The author will receive 100 free reprints (without covers) only if the page charge is honored. Detailed instructions will accompany the proofs.



IEEE

TRANSACTIONS ON

AUTOMATIC CONTROL



A PUBLICATION OF THE IEEE CONTROL SYSTEMS SOCIETY

FEBRUARY 1995

VOLUME 40

NUMBER 2

IETAC (ISSN 0018-9286)

Canning the Issue 197

Editorial — The 1994 George S. Axelby Outstanding Paper Award 199

PAPERS

Adaptive Control of Plants with Unknown Hystereses	<i>G. Tao and P. V. Kokotović</i>	200
Robust Stability Under a Class of Nonlinear Parametric Perturbations	<i>M. Fu, S. Dasgupta, and V. Blondel</i>	213
Discrete Time Observers for Singularly Perturbed Continuous Time Systems	<i>K. R. Shouse and D. G. Taylor</i>	224
Adaptive Back-Pressure Congestion Control Based on Local Information	<i>L. Tassiulas</i>	236
Stability of Queueing Networks and Scheduling Policies	<i>P. R. Kumar and S. P. Meyn</i>	251
Second-Order Properties of Families of Discrete Event Systems	<i>R. Rajan and R. Agrawal</i>	261

TECHNICAL NOTES AND CORRESPONDENCE

Regional Pole Placement of Multivariable Systems Under Control Structure Constraints	<i>S. S. Keerthi and M. S. Phatak</i>	272
Continuous Robust Control Design for Nonlinear Uncertain Systems Without <i>a Priori</i> Knowledge of Control Direction	<i>I. Kaloust and Z. Qu</i>	276
On the Ordering of Optimal Hedging Points in a Class of Manufacturing Flow Control Models	<i>G. Liberopoulos and J.-Q. Hu</i>	282
Recursive Identification Method for MISO Wiener-Hammerstein Model	<i>M. Boutayeb and M. Darouach</i>	287
Production Rate Control for Failure-Prone Production Systems With No Backlog Permitted	<i>J. Q. Hu</i>	291
Relative Stability of a Linear Time-Varying Process with First-Order Nonlinear Time-Varying Feedback	<i>J. S. Anas</i>	296
Properties of Optimal Weighted Sensitivity Designs	<i>K. E. Lenz</i>	298
A Periodic Fixed-Structure Approach to Multirate Control	<i>W. M. Haddad and V. Kapila</i>	301
Pursuing a Maneuvering Target Which Uses a Random Process for Its Control	<i>V. E. Beneš, K. L. Helmes, and R. W. Rishel</i>	307
A Subspace Fitting Method for Identification of Linear State-Space Models	<i>A. Swindlehurst, R. Roy, B. Ottersten, and T. Kailath</i>	311
Consistency of Modified LS Estimation Method for Identifying 2-D Noncausal SAR Model Parameters	<i>P. Y. Zhao and J. Litva</i>	316
A Robust Hybrid Stabilization Strategy for Equilibria	<i>J. Guckenheimer</i>	321
Adaptive Control of Systems with Unknown Output Backlash	<i>G. Tao and P. V. Kokotović</i>	326

(Continued on back cover)



The Control Systems Society is an organization within the framework of the IEEE of members with professional interest in automatic control. All members of the IEEE are eligible for membership in the Society and will receive the TRANSACTIONS upon payment of the annual Society membership fee of \$15.00. For further information write to the address below. Member copies of Transactions/Journals are for personal use only.

IEEE TRANSACTIONS ON AUTOMATIC CONTROL

Founding Editor: George S. Axelby

EDITORIAL BOARD

J. BAILEY *Editor in Chief*
Dep. Aerospace/Mechanical Eng.
Boston Univ.
110 Cummings St.
Boston, MA 02215
(617) 353-9848

C. CASSANDRAS *Editor Tech Notes and Corresp*
Dep. Electrical/Computer Eng.
Univ. Massachusetts
Amherst, MA 01003
(413) 545-0870

D. S. NAIIDU *Assoc. Editor Book Reviews*
College Eng.
Idaho State Univ.
833 South Eighth St.
Pocatello, ID 83209
(208) 236-2307

Associate Editors, Technical

M. ARAKI
Kyoto Univ.
A. BACH
Twente Univ. Technol.
A. M. BLOCH
Ohio State Univ.
J. A. BURNS
VPI & SU
C. CANUDAS DE WIL
ENSIEG INPC
E. K. P. CHONG
Purdue Univ.
M. DAHGH
Univ. Calif. Santa Barbara
M. DAHLH
Mass. Inst. Technol.
W. P. DAYAWANSA
Univ. Maryland

M. FI
Univ. Newcastle
W. B. GONG
Univ. Massachusetts
N. GUNDES
Univ. Calif. Davis
J. HAUSER
Univ. Colorado
C. V. HOLLOI
Univ. Massachusetts
G. HUANG
Texas A&M Univ.
F. JABBAR
Univ. Calif. Irvine
S. LAJOURNE
Univ. Michigan
J. LASERRI
CNRS

B. LUTMAN
Mississippi State
P. NAIN
INRIA
R. NIKOUKHAH
INRIA
K. PASSINO
Ohio State Univ.
J. B. POMET
INRIA
T. A. POSBURN
Univ. Minnesota
W. REN
Univ. Calif. Berkeley
A. SABERI
Washington State Univ.
J. SHAMMA
Univ. Texas, Austin

V. SOLO
Macquarie Univ.
A. A. STOKROVICH
Eindhoven Univ. Technol.
J. SUN
Ford Research Lab.
A. TESI
Univ. di Firenze
A. J. VAN DER SCHAF
Univ. Twente
P. VAN DOORN
Louvain Univ.
A. VINCIO
Universita di Siena
F. YU
Univ. Arkansas
G. YIN
Wayne State Univ.

Associate Editors at Large

A. BENVENISTO
IRISA/INRIA
P. CROUCH
Arizona State Univ.
D. I. ELIOT
Univ. Maryland

A. ISIDORI
Univ. Roma La Sapienza &
Washington Univ.
P. R. KUMAR
Univ. Illinois Urbana
I. LUNG
Linköping Univ.

A. N. MICHI
Univ. Notre Dame
S. K. MILLER
Mass. Inst. Technol.
M. MORARI
Calif. Inst. Technol.

M. K. SAIN
Univ. Notre Dame
I. D. SONTAG
Rutgers Univ.
A. TITS
Univ. of Maryland

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

Officers

JAMES T. CAIN *President*
WALLACE S. READ *President Elect*
CHARLES W. TURNER *Secretary*
V. THOMAS RIJINI *Treasurer*
KENNETH R. LAKIE *Vice President Educational Activities*

JOEL B. SNYDER *Vice President Professional Activities*
W. KENNETH DAWSON *Vice President Publication Activities*
VIVIAN K. BHARGAVA *Vice President Regional Activities*
L. G. KILNER *Vice President Standards Activities*
BRUCE EISENSTEIN *Vice President Technical Activities*

LEAH JONG (L. J.) TARN *Director, Division A—Systems and Control Division*

Executive Staff

THEODORE W. HISSNY, JR. *Executive Director*
RICHARD D. SCHWARTZ *Admin. General Manager*
PHILIP HALL *Staff Executive Publications*
FRANK R. MOORE *Staff Executive Volunteer Activities*

IRVING ENGELSON *Staff Director Corporate Activities*
W. R. HABINGREUTHER *Staff Director Customer Service Center*
PETER A. LEWIS *Staff Director Educational Activities*
MELVIN I. OLKEN *Staff Director Regional Activities*

ANDREW G. SALEM *Staff Director Standards Activities*
W. THOMAS SUTELLE *Staff Director Professional Activities*
ROBERT T. WANGELMANN *Staff Director Technical Activities*

Transactions/Journals Department

Director PATRICIA WALKER
Manager GAIL S. FIREN
Electronic Production Manager JERI L. UZZO
Managing Editor VALERIE CAMMARATA

Senior Editor GUERARDINI L. KROHN *Associate Editors* HELENE DORTCHIMER, DAWN SEETHI WHITE

IEEE TRANSACTIONS ON AUTOMATIC CONTROL is published monthly by The Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon IEEE, the Society/Council, or its members. **IEEE Corporate Office** 345 East 47th Street, New York, NY 10017 2394. **IEEE Operations Center** 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. **NJ Telephone** 908-981-0000. **Price/Publication Information** Individual copies: IEEE Members \$10.00 (first copy only); nonmembers \$20.00 per copy. (Note: Add \$4.00 postage and handling charge to any order from \$1.00 to \$50.00 including prepaid orders.) Member and nonmember subscription prices available on request. Available in microfiche and microfilm. **Copyright and Reprint Permissions** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons provided the per copy fee indicated at the bottom of the first page is paid through the Copyright Clearance Center, 29 Congress Street, Salem, MA 01970. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. Copyright ©1995 by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Second class postage paid at New York, NY, and at additional mailing offices. **Postmaster** Send address changes to IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

(Continued from front cover)

Reciprocal Processes on a Tree—Modeling and Estimation Issues.....	<i>R. W. Dijkerman, R. R. Mazumdar, and A. Bagchi</i>	330
Comments on the Loop Transfer Recovery	<i>A. Rachid</i>	335
Sensitivity Properties of Multirate Feedback Control Systems, Based on Eigenstructure Assignment	<i>R. J. Patton, G. P. Liu, and Y. Patel</i>	337
Optimality Conditions for Truncated Kautz Networks with Two Periodically Repeating Complex Conjugate Poles.....	<i>T. Oliveira e Silva</i>	342
Stable Adaptive Control of a Class of First-Order Nonlinearly Parameterized Plants	<i>J. D. Bošković</i>	347
Multiproduct Production/Inventory Control Under Random Demands	<i>J. Qiu and R. Loulou</i>	350
The Partial Model Matching or Partial Disturbance Rejection Problem: Geometric and Structural Solutions	<i>M. Malabre and J. C. Martinez Garcia</i>	356
Absolute Stability Criteria for Multiple Slope-Restricted Monotonic Nonlinearities.....	<i>W. M. Haddad and V. Kapila</i>	361
Simultaneous Disturbance Rejection and Regular Row by Row Decoupling With Stability: A Geometric Approach.....	<i>J. C. Martinez Garcia and M. Malabre</i>	365
Robust Controller Design for Delay Systems in the Gap-Metric	<i>A. Kojima and S. Ishijima</i>	370
A New Balanced Canonical Form for Stable Multivariable Systems.....	<i>B. Hanzon</i>	374
Boundary Fractional Derivative Control of the Wave Equation	<i>B. Mbodje and G. Montseny</i>	378
Observer-Based Parameter Identifiers for Nonlinear Systems with Parameter Dependencies	<i>S. Sheikholeslam</i>	382
Computation of Approximate Null Vectors of Sylvester and Lyapunov Operators	<i>A. R. Ghavimi and A. J. Laub</i>	387
ANNOUNCEMENTS		392

IEEE CONTROL SYSTEMS SOCIETY BOARD OF GOVERNORS

Executive Officers

D P AHERTON, President Univ Sussex Brighton, UK BN1 90T	M K MASTIN, President Elect Texas Instruments Plano, TX 75075	J D BIRDWELL, Secretary Administrator Univ Tennessee Knoxville, TN 37996	P J ANTSAKLIS, Vice President Conference Activities Univ Notre Dame Notre Dame, IN 46556
N H McCLAMROCK, Vice President Financial Affairs Univ Michigan Ann Arbor, MI 48109	D P LOOZE, Vice President Member Activities Univ Massachusetts Amherst MA 01003	A N MICHEL, Vice President Technical Affairs Univ Notre Dame Notre Dame, IN 46556	S YURKOVICH, Vice President Publication Activities Ohio State Univ Columbus, OH 43210
J BAILLIET, Editor-in Chief <i>Transactions on Automatic Control</i>	B H KROGH, Editor <i>Transactions on Control Systems Technology</i>	S YURKOVICH, Editor <i>IEEE Control Systems Magazine</i>	

Members

<i>Term Ending December 31 1995</i>	<i>Term Ending December 31 1995 (Appointed)</i>	<i>Term Ending December 31 1996</i>	<i>Term Ending December 31 1997</i>
J E ACKERMANN P E CAINE P K HOUP N H McCLAMROCK A N MICHEL M P POLIS	G DUMONT C SCHRAEDER M SPONG B SRIDHAR S VIJTAI RAO G GOODWIN	P J ANTSAKLIS K FURLA T I JOHNSON M K MASTIN I SHAW S YURKOVICH	A ANNASWAMY H KIMURA B KROGH K PASSINO D RIPPERGER M GIVERS

INFORMATION FOR AUTHORS

In the IEEE TRANSACTIONS ON AUTOMATIC CONTROL the IEEE Control Systems Society publishes high quality papers on the theory design and application of control systems. The TRANSACTIONS is published monthly.

Two types of contributions are regularly considered:

1) *Papers* - Presentation of significant research, development, or application of control concepts.

2) *Technical Notes and Correspondence* - Brief technical notes, comments on published areas or established control topics, corrections to papers and notes published in the TRANSACTIONS.

In addition, special papers (tutorials, surveys, and perspectives on the theory and applications of control systems topics) are solicited. Authors are urged to contact one of the Associate Editors at Large before submitting such papers.

Submission of a manuscript signifies that it has been neither copyrighted, published, nor submitted or accepted for publication elsewhere.

Submitted manuscripts must be typewritten in English. All submitted manuscripts should be as concise as possible. Technical Notes and Correspondence are normally limited to 12 double spaced typed pages. Papers of length exceeding 40 double spaced typed pages are strongly discouraged. The Editor reserves the right to refuse to consider such papers.

A. Process for Submission of Manuscript

- 1) *Papers* - Send seven copies of the paper together with two copies of a transmittal letter enclosed in one package to the Editor in Chief. These will be subject to a full review procedure, and a decision on whether or not to accept the paper will be made by members of the Transactions Editorial Board.
- 2) *Technical Notes and Correspondence* - Technical Notes and Correspondence should be sent to the Editor for Technical Notes and Correspondence. Five copies are required. Decisions are made on the basis of a simplified review procedure.
- 3) Original illustrations should not be sent until requested, but authors should be ready to submit these immediately upon acceptance for publication.
- 4) Authors of accepted Papers are requested to supply their biographies (100 words or less) and photographs. For style, see biographies in this issue.
- 5) Enclose two copies of your letter of transmittal. Do not send letters of transmittal under separate cover. Give your preferred address for correspondence. Inform the Editor-in-Chief or Editor for Technical Notes and Correspondence of any change of address.
- 6) Authors of accepted manuscripts will be required to provide the final version of their manuscript on a computer diskette along with the hard copy.

B. Copyright

It is the policy of the IEEE to own the copyright to the technical contribution it publishes on behalf of the interests of the IEEE, its authors, and their employers, and to facilitate the appropriate reuse of this material by others. To comply with the U.S. Copyright Law, authors are required to sign an IEEE copyright form before publication. This form, a copy of which appears in the January 1995 issue of this journal, returns to authors and their employers full rights to reuse their material for their own purposes. Authors must submit a signed copy of this form with their manuscripts.

C. Style for Manuscript

- 1) First page must contain: a) Title of paper (without symbols); b) Author(s) and affiliation(s); c) Abstract (not exceeding 300 words for Papers or 75 words for Technical Notes and Correspondence, and without equations, references, or footnotes); d) Complete mailing address (telephone number and, if available, electronic mail (email) address and facsimile (fax) number of each author); e) Preferred address for correspondence and return of proofs; and f) Footnotes (if desired) containing acknowledgment of financial or other support.
- 2) If possible, please prepare copies using two sides of each page. Authors should be aware, however, that after acceptance of a manuscript, a one-sided copy must be provided for IEEE production purposes.
- 3) Provide an Introduction that includes a statement of the purpose and contribution of the paper.
- 4) If appropriate, indicate advantages, limitations, and possible applications in a Conclusion section.
- 5) References should be numbered and appear in a separate bibliography at the end of the paper. Use numerals in square brackets to cite references, e.g., [15]. References should be complete and in IEEE style (see examples in this issue).

D. Style for Illustrations

- 1) It is in the author's interest to submit professional quality illustrations. Drafting or art service cannot be provided by the IEEE.
- 2) Original drawings should be in black ink on white background. Maximum size is restricted to 21.6 by 27.9 cm. Glossy prints of illustrations are also acceptable.
- 3) All lettering should be large enough to permit legible reduction of the figure to column width, sometimes as small as one quarter of the original size. Typed lettering is usually not acceptable on figures.
- 4) Lightly pencil each figure number on the back of each original illustration. Captions should not appear on figures.
- 5) Provide a separate sheet listing all figure captions in proper style for the typesetter, e.g., Fig. 5. The error variance for the optimal filter.
- 6) Contributors' photographs should measure between 1.6 cm and 9.5 cm across the widest part of the head. The overall size of the photographic paper used can be anything from passport size to 27.9 cm.

E. Mandatory Overlength Page Charges

A mandatory page charge is imposed on all papers exceeding eight TRANSACTIONS pages (about 28 double-spaced typed pages) in length, including illustrations. The charge is \$125 per page for each page over the first eight and is a prerequisite for publication. Details are provided at the time of acceptance; authors are, however, urged to keep this in mind when submitting and revising their papers.

F. Page Charges

After a manuscript has been accepted for publication, the author's company or institution will be approached with a request to pay a charge of \$110 per page to cover part of the cost of publication. Payment of page charges for this IEEE TRANSACTIONS-like journals of other professional societies is not a necessary prerequisite for publication. The author will receive 100 free reprints (without covers) only if the page charge is honored. Detailed instructions will accompany the proofs.



IEEE

TRANSACTIONS ON

AUTOMATIC CONTROL



A PUBLICATION OF THE IEEE CONTROL SYSTEMS SOCIETY

MARCH 1995

VOLUME 40

NUMBER 3

IETAA9

(ISSN 0018-9286)

Scanning the Issue	393
--------------------------	-----

PAPERS

Observer Design for Nonlinear Systems with Discrete-Time Measurements	<i>P. E. Moraal and J. W. Grizzle</i>	395
A Probabilistic Approach to Multivariable Robust Filtering and Open-Loop Control	<i>K. Öhrn, A. Ahlén, and M. Sternad</i>	405
A New Model for Control of Systems with Friction	<i>C. Canudas de Wit, H. Olsson, K. J. Åström, and P. Lischinsky</i>	419
Adaptive Nonlinear Design with Controller-Identifier Separation and Swapping	<i>M. Kistić and P. V. Kokotović</i>	426
Approximate Decoupling and Asymptotic Tracking for MIMO Systems	<i>D. N. Godbole, and S. S. Sastry</i>	441
A Generalized Orthonormal Basis for Linear Dynamical Systems	<i>P. S. C. Heuberger, P. M. J. Van den Hof, and O. H. Bosgra</i>	451
H_∞ Control via Measurement Feedback for General Nonlinear Systems	<i>A. Isidori and W. Kang</i>	466

TECHNICAL NOTES AND CORRESPONDENCE

Perturbation Bounds for Root-Clustering of Linear Systems in a Specified Second Order Subregion	<i>W. Bakker, J. S. Luo, and A. Johnson</i>	473
Comments on "Strictly Positive Real Transfer Functions Revisited"	<i>H. J. Marquez and C. J. Damaren</i>	478
On Interval Polynomials with No Zeros in the Unit Disc	<i>V. Blondel</i>	479
The Logical Control of an Elevator	<i>D. N. Dyck and P. E. Caines</i>	480
Robust Stability Criteria for Dynamical Systems Including Delayed Perturbations	<i>H. Wu and K. Mizukami</i>	487
An Efficient Method for Unconstrained Optimization Problems of Nonlinear Large Mesh-Interconnected Systems	<i>S.-Y. Lin and C.-H. Lin</i>	490
On the Possible Divergence of the Projection Algorithm	<i>E. Lefeber and J. W. Polderman</i>	495
A Comment on the Method of the Closest Unstable Equilibrium Point in Nonlinear Stability Analysis	<i>E. Noldus and M. Loccupier</i>	497
Boundaries of Conditional Quadratic Forms—A Comment on "Stabilization via Static Output Feedback"	<i>D. Cheng and C. F. Martin</i>	500

(Continued on back cover)



The Control Systems Society is an organization, within the framework of the IEEE, of members with professional interest in automatic control. All members of the IEEE are eligible for membership in the Society and will receive the *TRANSACTIONS* upon payment of the annual Society membership fee of \$15.00. For further information, write to the address below. *Member copies of Transactions/Journals are for personal use only.*

IEEE TRANSACTIONS ON AUTOMATIC CONTROL

Founding Editor: George S. Axelby

EDITORIAL BOARD

J BAILLIEUL, *Editor-in-Chief*
Dep. Aerospace/Mechanical Eng.
Boston Univ
110 Cummings St
Boston, MA 02215
(617) 353-9848

C CASSANDRAS, *Editor, Tech Notes and Corresp*
Dep. Electrical/Computer Eng
Univ. Massachusetts
Amherst, MA 01003
(413) 545-0870

D S NAIDU, *Assoc. Editor, Book Reviews*
College Eng
Idaho State Univ
833 South Eighth St
Pocatello, ID 83209
(208) 236-2307

Associate Editors, Technical

M ARAKI
Kyoto Univ.
A BAOCHI
Twente Univ Technol

M DAHLEH
Mass Inst Technol
W P DAYAWANSA
Univ Maryland

G HUANO
Texas A&M Univ
F JABBARI
Univ Calif Irvine

K PASSINO
Ohio State Univ
J-B POMET
INRIA

A A STOOORVOGEL
Eindhoven Univ Technol
J SUN
Ford Research Lab

A M BLOCH
Ohio State Univ

M FU
Univ Newcastle

S LAPORTUNE
Univ Michigan

T A POSBERGII
Univ Minnesota

A TESTI
Univ di Firenze

J A BURNS
VPI & SU

W-B GONG
Univ Massachusetts

J LASSERRE
CNRS

W REN
Univ Calif Berkeley

A J VAN DER SCHAFT
Univ Twente

C. CANUDAS DE WIT
ENSIEG-INPG

N GUNDES
Univ Calif -Davis

B LEHMAN
Mississippi State

A SABERI
Washington State Univ

P VAN DOOREN
Louvain Univ

E K P CHONG
Purdue Univ

J HAUSER
Univ Colorado

P NAIN
INRIA

J SHAMMA
Univ Texas Austin

A VICINO
Universita di Siena

M DAHLEH
Univ Calif -Santa Barbara

C V HOLLOT
Univ Massachusetts

R NIKOUKHAH
INRIA

V SOLO
Macquarie Univ

E YAZ
Univ Arkansas

G YIN
Wayne State Univ

Associate Editors at Large

A BENVENISTE
IRISA/INRIA

A ISIDORI
Univ Roma La Sapienza &
Washington Univ

L LJUNG
Linkoping Univ

S K MITTER
Mass Inst Technol

M K SAIN
Univ Notre Dame

P CROUCH
Arizona State Univ

P R KUMAR
Univ Illinois Urbana

A N MICHEL
Univ Notre Dame

M MORARI
Calif Inst Technol

E D SONTAG
Rutgers Univ

D L ELLIOTT
Univ Maryland

A TITS
Univ of Maryland

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC

Officers

JAMES T. CAIN, *President*
WALLACE S. READ, *President-Elect*
CHARLES W. TURNER, *Secretary*
V. THOMAS RHYNE, *Treasurer*
KENNETH R. LAKPR, *Vice President Educational Activities*

JOEL B. SNYDER, *Vice President, Professional Activities*
W. KENNETH DAWSON, *Vice President Publication Activities*
VIJAY K. BHARGAVA, *Vice President Regional Activities*
E. G. KIENER, *Vice President Standards Activities*
BRUCE EISENSTEIN, *Vice President Technical Activities*

TZYH-JONG TARN, *Director, Division X -Systems and Control Division*

Executive Staff

THEODORE W. HISSEY, JR., *Executive Director*
RICHARD D. SCHWARTZ, *Acting General Manager*
PHYLLIS HALL, *Staff Executive, Publications*
FRANK R. MOORE, *Staff Executive, Volunteer Activities*

IRVING ENGELSON, *Staff Director, Corporate Activities*
PETER A. LEWIS, *Staff Director, Educational Activities*
MELVIN I. OLKEN, *Staff Director, Regional Activities*

ANDREW G. SALEM, *Staff Director, Standards Activities*
W. THOMAS SUTTLE, *Staff Director, Professional Activities*
ROBERT T. WANGEMANN, *Staff Director, Technical Activities*

Transactions/Journals Department

Director PATRICIA WALKER
Manager GAIL S. FERENC
Electronic Production Manager JERI L. UZZO
Managing Editor VALERIE CAMMARATA

Senior Editor GERALDINE E. KROIN *Associate Editor* DAWN SPETH WHITE *Assistant Editor* MICHELLE MEEH

IEEE TRANSACTIONS ON AUTOMATIC CONTROL is published monthly by The Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon IEEE, the Society/Council or its members. IEEE Corporate Office: 345 East 47th Street, New York, NY 10017-2394. IEEE Operations Center: 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. NJ Telephone: 908-981-0060. Price/Publication Information: Individual copies: IEEE Members \$10.00 (first copy only), nonmembers \$20.00 per copy. (Note: Add \$4.00 postage and handling charge to any order from \$1.00 to \$50.00, including prepaid orders.) Member and nonmember subscription prices available on request. Available in microfiche and microfilm. Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee indicated at the bottom of the first page is paid through the Copyright Clearance Center, 29 Congress Street, Salem, MA 01970. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. Copyright ©1995 by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Second class postage paid at New York, NY and at additional mailing offices. Postmaster: Send address changes to IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. Printed in U.S.A.

(Continued from front cover)

On Robust Stability of 2-D Discrete Systems	<i>W.-S. Lu</i>	502
Multivariable System Identification via Continued-Fraction Approximation	<i>R. Johansson</i>	507
All Fixed-Order H_∞ Controllers: Observer-Based Structure and Covariance Bounds	<i>T. Iwasaki and R. E. Skelton</i>	512
Intrinsic Difficulties in Using the Doubly-Infinite Time Axis for Input–Output Control Theory	<i>T. T. Georgiou and M. C. Smith</i>	516
Regional Observability of a Thermal Process	<i>A. El Jai, E. Zerrik, M. C. Simon, and M. Amouroux</i>	518
Towards a Generalized Regulation Scheme for Oscillatory Systems via Coupling Effects	<i>K. L. Tuer, M. F. Golnaraghi, and D. Wang</i>	522
μ -Synthesis of an Electromagnetic Suspension System	<i>M. Fujita, T. Namerikawa, F. Matsumura, and K. Uchida</i>	530
Parameter-Dependent Lyapunov Functions and the Popov Criterion in Robust Analysis and Synthesis	<i>W. M. Haddad and D. S. Bernstein</i>	536
Design of L-Q Regulators for State Constrained Continuous-Time Systems	<i>C. E. T. Dórea and B. E. A. Milani</i>	544
The Finite Inclusions Theorem	<i>R. D. Kaminsky and T. E. Djaferis</i>	549
Further Results on Rational Approximations of \mathcal{L}^1 Optimal Controllers	<i>Z.-Q. Wang, M. Sznajer, and F. Blanchini</i>	552
Supervisory Control of Timed Discrete-Event Systems under Partial Observation	<i>F. Lin and W. M. Wonham</i>	558
Apology and Correction to “Process Control and Machine Learning: Rule-Based Incremental Control”	<i>D. Luzeaux and B. Zavidovique</i>	562

BOOK REVIEWS

Mathematical Control Theory: Deterministic Finite-Dimensional Systems—Eduardo D. Sontag	<i>Reviewed by S. P. Boyd</i>	563
---	-------------------------------	-----

REVISED 1994 INDEX	Follows page	564
--------------------------	--------------	-----

IEEE CONTROL SYSTEMS SOCIETY

BOARD OF GOVERNORS

Executive Officers

D. P. ATHERTON, President
Univ. Sussex
Brighton, UK BN1 90T

M. K. MASTEN, President-Elect
Texas Instruments
Plano, TX 75075

J. D. BIRDWELL, Secretary-Administrator
Univ. Tennessee
Knoxville, TN 37996

P. J. ANTSAKLIS, Vice President
Conference Activities
Univ. Notre Dame
Notre Dame, IN 46556

N. H. McCLAMROCH, Vice President, Financial Affairs
Univ. Michigan
Ann Arbor, MI 48109

D. P. LOOZE, Vice President, Member Activities
Univ. Massachusetts
Amherst, MA 01003

A. N. MICHEL, Vice President, Technical Affairs
Univ. Notre Dame
Notre Dame, IN 46556

S. YURKOVICH, Vice President, Publication Activities
Ohio State Univ.
Columbus, OH 43210

J. BAILLIEUL, Editor-in-Chief
Transactions on Automatic Control

B. H. KROGH, Editor
Transactions on Control Systems Technology

S. YURKOVICH, Editor
IEEE Control Systems Magazine

Members

Term Ending December 31, 1995

*Term Ending December 31, 1995
(Appointed)*

Term Ending December 31, 1996

Term Ending December 31, 1997

J. E. ACKERMANN
P. E. CAINES
P. K. HOUP
N. H. McCLAMROCH
A. N. MICHEL
M. P. POLIS

G. DUMONT
C. SCHRADER
M. SPONG
B. SRIDHAR
S. VITTAL RAO
G. GOODWIN

P. J. ANTSAKLIS
K. FURUTA
T. L. JOHNSON
M. K. MASTEN
L. SHAW
S. YURKOVICH

A. ANNASWAMY
H. KIMURA
B. KROGH
K. PASSINO
D. REPPERGER
M. GEVERS

INFORMATION FOR AUTHORS

In the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, the IEEE Control Systems Society publishes high-quality papers on the theory, design, and application of control systems. The TRANSACTIONS is published monthly.

Two types of contributions are regularly considered:

- 1) *Papers*—Presentation of significant research, development, or application of control concepts.
- 2) *Technical Notes and Correspondence*—Brief technical notes, comments on published areas or established control topics, corrections to papers and notes published in the TRANSACTIONS

In addition, special papers (tutorials, surveys, and perspectives on the theory and applications of control systems topics) are solicited. Authors are urged to contact one of the Associate Editors at large before submitting such papers.

Submission of a manuscript signifies that it has been neither copyrighted, published, nor submitted or accepted for publication elsewhere.

Submitted manuscripts must be typewritten in *English*. All submitted manuscripts should be as concise as possible. Technical Notes and Correspondence are normally limited to 12 double-spaced, typed pages. Papers of length exceeding 40 double-spaced, typed pages are strongly discouraged. The Editor reserves the right to refuse to consider such papers.

A. Process for Submission of Manuscript

- 1) *Papers*. Send seven copies of the paper together with two copies of a transmittal letter enclosed in one package to the Editor-in-Chief. These will be subject to a full review procedure, and a decision on whether or not to accept the paper will be made by members of the Transactions Editorial Board.
- 2) *Technical Notes and Correspondence*: Technical Notes and Correspondence should be sent to the Editor for Technical Notes and Correspondence; five copies are required. Decisions are made on the basis of a simplified review procedure.
- 3) Original illustrations should not be sent until requested, but authors should be ready to submit these immediately upon acceptance for publication.
- 4) Authors of accepted Papers are requested to supply their biographies (100 words or less) and photographs. For style, see biographies in this issue.
- 5) Enclose two copies of your letter of transmittal. Do not send letters of transmittal under separate cover. Give your preferred address for correspondence. Inform the Editor-in-Chief or Editor for Technical Notes and Correspondence of any change of address.
- 6) Authors of accepted manuscripts will be required to provide the final version of their manuscript on a computer diskette along with the hard copy.

B. Copyright

It is the policy of the IEEE to own the copyright to the technical contribution it publishes on behalf of the interests of the IEEE, its authors and their employers, and to facilitate the appropriate reuse of this material by others. To comply with the U.S. Copyright Law, authors are required to sign an IEEE copyright form before publication. This form, a copy of which appears in the January 1995 issue of this journal, returns to authors and their employers full rights to reuse their material for their own purposes. Authors must submit a signed copy of this form with their manuscripts.

C. Style for Manuscript

- 1) First page must contain: a) Title of paper (without symbols), b) Author(s) and affiliation(s), c) Abstract (not exceeding 300 words for Papers or 75 words for Technical Notes and Correspondence, and without equations, references, or footnotes), d) Complete mailing address, telephone number, and, if available, electronic mail (email) address and facsimile (fax) number of each author, e) Preferred address for correspondence and return of proofs, and f) Footnotes (if desired) containing acknowledgment of financial or other support.
- 2) If possible, please prepare copies using two sides of each page. Authors should be aware, however, that after acceptance of a manuscript, a one-sided copy must be provided for IEEE production purposes.
- 3) Provide an Introduction that includes a statement of the purpose and contribution of the paper.
- 4) If appropriate, indicate advantages, limitations, and possible applications in a Conclusion section.
- 5) References should be numbered and appear in a separate bibliography at the end of the paper. Use numerals in square brackets to cite references, e.g., [15]. References should be complete and in IEEE style (see examples in this issue).

D. Style for Illustrations

- 1) It is in the author's interest to submit professional quality illustrations. Drafting or art service cannot be provided by the IEEE.
- 2) Original drawings should be in black ink on white background. Maximum size is restricted to 21.6 by 27.9 cm. Glossy prints of illustrations are also acceptable.
- 3) All lettering should be large enough to permit legible reduction of the figure to column width, sometimes as small as one quarter of the original size. Typed lettering is usually not acceptable on figures.
- 4) Lightly pencil each figure number on the back of each original illustration. Captions should not appear on figures.
- 5) Provide a separate sheet listing all figure captions, in proper style for the typesetter, e.g., "Fig. 5. The error variance for the optimal filter."
- 6) Contributors' photographs should measure between 1.6 cm and 9.5 cm across the widest part of the head. The overall size of the photographic paper used can be anything from passport size to 27.9 cm.

E. Mandatory Overlength Page Charges

A mandatory page charge is imposed on all papers exceeding eight TRANSACTIONS pages (about 28 double-spaced, typed pages) in length, including illustrations. The charge is \$125 per page for each page over the first eight and is a prerequisite for publication. Details are provided at the time of acceptance; authors are, however, urged to keep this in mind when submitting and revising their papers.

F. Page Charges

After a manuscript has been accepted for publication, the author's company or institution will be approached with a request to pay a charge of \$110 per page to cover part of the cost of publication. Payment of page charges for this IEEE TRANSACTIONS, like journals of other professional societies, is not a necessary prerequisite for publication. The author will receive 100 free reprints (without covers) only if the page charge is honored. Detailed instructions will accompany the proofs.

**IEEE****TRANSACTIONS ON**

AUTOMATIC CONTROL

**A PUBLICATION OF THE IEEE CONTROL SYSTEMS SOCIETY****APRIL 1995****VOLUME 40****NUMBER 4****IETAA9****(ISSN 0018-9286)**

✦ Scanning the Issue	601
----------------------------	-----

PAPERS

Adjoint and Hamiltonian Input-Output Differential Equations	603
..... <i>P. E. Crouch, F. Lannabhi-Lagarigue, and A. J. van der Schaft</i>	
Language Convergence in Controlled Discrete-Event Systems	616
..... <i>Y. M. Willner and M. Heymann</i>	
Concurrent Vector Discrete-Event Systems	628
..... <i>Y. Li and W. M. Wonham</i>	
Structure of Model Uncertainty for a Weakly Corrupted Plant	639
..... <i>T. Zhou and H. Kimura</i>	
Explicit Formulas for Optimally Robust Controllers for Delay Systems.....	656
..... <i>H. Dym, T. T. Georgiou, and M. C. Smith</i>	
Stochastic System Identification with Noisy Input-Output Measurements Using Polyspectra.....	670
..... <i>J. K. Tugnait and Y. Ye</i>	
Minimum Bias Priors for Estimating Parameters of Additive Terms in State-Space Models	684
..... <i>B. Hochwald and A. Nehorai</i>	

TECHNICAL NOTES AND CORRESPONDENCE

Nonlinear L_1 Optimal Controllers for Linear Systems.....	694
..... <i>A. A. Stoorvogel</i>	
Simultaneous Observation of Linear Systems	696
..... <i>Y. X. Yao, M. Darouach, and J. Schaefer</i>	
An Algorithm for Computing the Mask Value of the Supremal Normal Sublanguage of a Legal Language	699
..... <i>M. Barbeau, G. Custeau, and R. St-Denis</i>	
Decentralized Robust Control of Uncertain Interconnected Systems with Prescribed Degree of Exponential Convergence	704
..... <i>Z. Gong</i>	
Hedging-Point Production Control with Multiple Failure Modes	707
..... <i>P. Glasserman</i>	
A Globally Optimal Minimax Solution for Spectral Overbounding and Factorization.....	712
..... <i>R. E. Scheid and D. S. Bayard</i>	
A Lower Bound for Limiting Time Delay for Closed-Loop Stability of an Arbitrary SISO Plan.....	717
..... <i>R. Devanathan</i>	
A New Reduction Technique for a Class of Singularly Perturbed Optimal Control Problems	721
..... <i>V. Gaitsgory</i>	
Pole Assignment with Robust Stability.....	725
..... <i>M. E. Halpern, R. J. Evans, and R. D. Hill</i>	
The Carathéodory-Fejér Problem and $\mathcal{H}_\infty/\ell_1$ Identification: A Time Domain Approach	729
..... <i>J. Chen and C. N. Nett</i>	
Pole Assignment for Linear Periodic Systems by Memoryless Output Feedback	735
..... <i>D. Aeyels and J. L. Willems</i>	
Fault Detection and Isolation for Unstable Linear Systems.....	740
..... <i>M. Kinnaert, R. Hanus, and Ph. Arte</i>	

(Continued on back cover)



The Control Systems Society is an organization, within the framework of the IEEE, of members with professional interest in automatic control. All members of the IEEE are eligible for membership in the Society and will receive the *TRANSACTIONS* upon payment of the annual Society membership fee of \$15.00. For further information, write to the address below. *Member copies of Transactions/Journals are for personal use only.*

IEEE TRANSACTIONS ON AUTOMATIC CONTROL

Founding Editor: George S. Axelby

EDITORIAL BOARD

J. BAILLIEUL, *Editor-in-Chief*
Dep. Aerospace/Mechanical Eng.
Boston Univ.
110 Cummings St.
Boston, MA 02215
(617) 353-9848

C. CASSANDRAS, *Editor, Tech Notes and Corresp*
Dep. Electrical/Computer Eng.
Univ. Massachusetts
Amherst, MA 01003
(413) 545-0870

D. S. NAIDU, *Assoc. Editor Book Reviews*
College Eng.
Idaho State Univ.
833 South Eighth St.
Pocatello, ID 83209
(208) 236-2307

Associate Editors, Technical

M. ARAKI
Kyoto Univ.
A. BAGCHI
Twente Univ. Technol.

M. DAHLEH
Mass. Inst. Technol.
W. P. DAYAWANSA
Univ. Maryland

G. HUANG
Texas A&M Univ.
F. JABBAR
Univ. Calif. Irvine

K. PASSINO
Ohio State Univ.
J.-B. POMET
INRIA

A. A. STOORVOGEL
Eindhoven Univ. Technol.
J. SUN
Ford Research Lab.

A. M. BLOCH
Ohio State Univ.

M. FU
Univ. Newcastle

S. LAFORTUNE
Univ. Michigan

T. A. POSBERGH
Univ. Minnesota

A. TESI
Univ. di Firenze

J. A. BURNS
VPI & SU

W.-B. GONG
Univ. Massachusetts

J. LASERRER
CNRS

W. REN
Univ. Calif. Berkeley

A. J. VAN DER SCHAFT
Univ. Twente

C. CANUDAS DE WIT
ENSIEG-INPG

N. GUNDEL
Univ. Calif.-Davis

B. LEHMAN
Mississippi State

A. SABERI
Washington State Univ.

P. VAN DOOREN
Louvain Univ.

E. K. P. CHONG
Purdue Univ.

J. HAUSER
Univ. Colorado

P. NAIN
INRIA

J. SHAMMA
Univ. Texas, Austin

A. VICINO
Universita di Siena

M. DAHLEH
Univ. Calif.-Santa Barbara

C. V. HOLLOT
Univ. Massachusetts

R. NIKOLKHAH
INRIA

V. SOLO
Macquarie Univ.

F. YAZ
Univ. Arkansas

G. YIN
Wayne State Univ.

Associate Editors at Large

A. BENVENISTE
IRISA/INRIA

A. ISIDORI
Univ. Roma La Sapienza &
Washington Univ.

L. LUNG
Linköping Univ.

S. K. MITTER
Mass. Inst. Technol.

M. K. SAIN
Univ. Notre Dame

P. CROUCH
Arizona State Univ.

P. R. KUMAR
Univ. Illinois Urbana

A. N. MICHEL
Univ. Notre Dame

M. MORARI
Calif. Inst. Technol.

F. D. SONTAG
Rutgers Univ.

D. L. ELLIOTT
Univ. Maryland

A. TITS
Univ. of Maryland

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

Officers

JAMES T. CAIN, *President*
WALLACE S. READ, *President Elect*
CHARLES W. TURNER, *Secretary*
V. THOMAS RHYNE, *Treasurer*
KENNETH R. LAKER, *Vice President Educational Activities*

JOEL B. SNYDER, *Vice President Professional Activities*
W. KENNETH DAWSON, *Vice President Publication Activities*
VIJAY K. BHARGAVA, *Vice President Regional Activities*
E. G. KIENER, *Vice President Standards Activities*
BRUCE EISENSTEIN, *Vice President Technical Activities*

TEYH-JONG FARN, *Director Division X—Systems and Control Division*

Executive Staff

THEODORI W. HISSEY, JR., *Executive Director*
RICHARD D. SCHWARTZ, *Acting General Manager*
PHYLLIS HALL, *Staff Executive Publications*
FRANK R. MOORE, *Staff Executive Volunteer Activities*

IRVING ENGLISH, *Staff Director Corporate Activities*
PETER A. LEWIS, *Staff Director Educational Activities*
MILVIN I. OLKEN, *Staff Director Regional Activities*

ANDREW G. SALEM, *Staff Director Standards Activities*
W. THOMAS SUTTLE, *Staff Director Professional Activities*
ROBERT T. WANGEMANN, *Staff Director Technical Activities*

Transactions/Journals Department

Director PATRICIA WALKER
Manager GAIL S. FRENC
Electronic Production Manager JEFF L. UZZO
Managing Editor VALERIE CAMMARATA

Senior Editor GERALDINE E. KROLIN *Associate Editor* DAWN SPETH WHITE *Assistant Editor* MICHELLE MEEH

IEEE TRANSACTIONS ON AUTOMATIC CONTROL is published monthly by The Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon IEEE, the Society/Council or its members. IEEE Corporate Office: 345 East 47th Street, New York, NY 10017-2394. IEEE Operations Center: 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. NJ Telephone: 908-981-0060. Price/Publication Information: Individual copies: IEEE Members \$10.00 (first copy only); nonmembers \$20.00 per copy. (Note: Add \$4.00 postage and handling charge to any order from \$1.00 to \$50.00, including prepaid orders.) Member and nonmember subscription prices available on request. Available in microfiche and microfilm. Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per copy fee indicated at the bottom of the first page is paid through the Copyright Clearance Center, 29 Congress Street, Salem, MA 01970. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. Copyright ©1995 by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved. Second class postage paid at New York, NY and at additional mailing offices. Postmaster: Send address changes to IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. Printed in U.S.A.

(Continued from front cover)

Zeros of Discretized Continuous Systems Expressed in the Euler Operator—An Asymptotic Analysis	<i>A. Tesfaye and M. Tomizuka</i>	743
Further Theoretical Results on Direct Strain Feedback Control of Flexible Robot Arms	<i>Z.-H. Luo and B. Guo</i>	747
H^∞ Optimal and Suboptimal Controllers for Infinite Dimensional SISO Plants	<i>O. Toker and H. Özbay</i>	751
Least Squares Type Algorithms for Identification in the Presence of Modeling Uncertainty	<i>E.-W. Bai and K. M. Nagpal</i>	756
Square-Root Bryson–Frazier Smoothing Algorithms	<i>P.G. Park and T. Kailath</i>	761
Robust H_∞ Control of Uncertain Nonlinear System via State Feedback	<i>T. Shen and K. Tamura</i>	766
Optimal Nonparametric Identification from Arbitrary Corrupt Finite Time Series	<i>J. Chen, C. N. Nett, and M. K. H. Fan</i>	769
Input Saturation and Global Stabilization of Nonlinear Systems via State and Output Feedback	<i>W. Lin</i>	776
Optimal Control for Systems with Deterministic Production Cycles	<i>J.-Q. Hu and D. Xiang</i>	782

BOOK REVIEWS

Analysis and Control of Nonlinear Infinite Dimensional Systems—V. Barbu	<i>Reviewed by H. O. Fattorini</i>	787
---	------------------------------------	-----

IEEE CONTROL SYSTEMS SOCIETY

BOARD OF GOVERNORS

Executive Officers

D. P. ATHERTON, President
Univ. Sussex
Brighton, UK BN1 90T

M. K. MASTEN, President-Elect
Texas Instruments
Plano, TX 75075

J. D. BIRDWELL, Secretary-Administrator
Univ. Tennessee
Knoxville, TN 37996

P. J. ANTSAKLIS, Vice President
Conference Activities
Univ. Notre Dame
Notre Dame, IN 46556

N. H. McCLAMROCH, Vice President,
Financial Affairs
Univ. Michigan
Ann Arbor, MI 48109

D. P. LOOZE, Vice President,
Member Activities
Univ. Massachusetts
Amherst, MA 01003

A. N. MICHEL, Vice President,
Technical Affairs
Univ. Notre Dame
Notre Dame, IN 46556

S. YURKOVICH, Vice President,
Publication Activities
Ohio State Univ.
Columbus, OH 43210

J. BAILLIEUL, Editor-in-Chief
Transactions on Automatic Control

B. H. KROGH, Editor
Transactions on Control Systems Technology

S. YURKOVICH, Editor
IEEE Control Systems Magazine

Members

Term Ending December 31, 1995

Term Ending December 31, 1995
(Appointed)

Term Ending December 31, 1996

Term Ending December 31, 1997

J. E. ACKERMANN
P. E. CAINES
P. K. HOUFF
N. H. McCLAMROCH
A. N. MICHEL
M. P. POLIS

G. DUMONT
C. SCHRADER
M. SPONG
B. SRIDHAR
S. VITTAL RAO
G. GOODWIN

P. J. ANTSAKLIS
K. FURUTA
T. L. JOHNSON
M. K. MASTEN
L. SHAW
S. YURKOVICH

A. ANNASWAMY
H. KIMURA
B. KROGH
K. PASSINO
D. REPPERGER
M. GEVERS

INFORMATION FOR AUTHORS

In the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, the IEEE Control Systems Society publishes high-quality papers on the theory, design, and application of control systems. The TRANSACTIONS is published monthly.

Two types of contributions are regularly considered:

1) *Papers*—Presentation of significant research, development, or application of control concepts.

2) *Technical Notes and Correspondence*—Brief technical notes, comments on published areas or established control topics, corrections to papers, and notes published in the TRANSACTIONS.

In addition, special papers (tutorials, surveys, and perspectives on the theory and applications of control systems topics) are solicited. Authors are urged to contact one of the Associate Editors at Large before submitting such papers.

Submission of a manuscript signifies that it has been neither copyrighted, published, nor submitted or accepted for publication elsewhere.

Submitted manuscripts must be typewritten in English. All submitted manuscripts should be as concise as possible. Technical Notes and Correspondence are normally limited to 12 double-spaced, typed pages. Papers of length exceeding 40 double-spaced, typed pages are strongly discouraged. The Editor reserves the right to refuse to consider such papers.

A. Process for Submission of Manuscript

- 1) *Papers*: Send seven copies of the paper together with two copies of a transmittal letter enclosed in one package to the Editor-in-Chief. These will be subject to a full review procedure, and a decision on whether or not to accept the paper will be made by members of the Transactions Editorial Board.
- 2) *Technical Notes and Correspondence*: Technical Notes and Correspondence should be sent to the Editor for Technical Notes and Correspondence; five copies are required. Decisions are made on the basis of a simplified review procedure.
- 3) Original illustrations should not be sent until requested, but authors should be ready to submit these immediately upon acceptance for publication.
- 4) Authors of accepted Papers are requested to supply their biographies (100 words or less) and photographs. For style, see biographies in this issue.
- 5) Enclose two copies of the letter of transmittal. Do not send letters of transmittal under separate cover. Give the authors preferred address for correspondence. Inform the Editor-in-Chief or Editor for Technical Notes and Correspondence of any change of address.
- 6) Authors of accepted manuscripts will be required to provide the final version of their manuscript on a computer disk in ASCII text file, along with two paper copies, or a final version can be emailed to "t-ac@ieee.org."

B. Copyright

It is the policy of the IEEE to own the copyright to the technical contributions it publishes. To comply with IEEE copyright policy, authors are required to sign an IEEE Copyright Transfer Form before publication in either the print or electronic medium. The form is provided upon approval of their manuscript. Authors must submit a signed copy of this form with their manuscripts.

C. Style for Manuscript

- 1) First page must contain: a) Title of paper (without symbols), b) Author(s) and affiliation(s), c) Abstract (not exceeding 300 words for Papers or 75 words for Technical Notes and Correspondence, and without equations, references, or footnotes), d) Complete mailing address, telephone number, and, if available, electronic mail (email) address and facsimile (fax) number of each author, e) Preferred address for correspondence and return of proofs, and f) Footnotes (if desired) containing acknowledgment of financial or other support.
- 2) Provide an Introduction that includes a statement of the purpose and contribution of the paper.
- 3) If appropriate, indicate advantages, limitations, and possible applications in a Conclusion section.
- 4) References should be numbered and appear in a separate bibliography at the end of the paper. Use numerals in square brackets to cite references, e.g., [15]. References should be complete and in IEEE style (see examples in this issue).

D. Style for Illustrations

Illustrations should be kept to a minimum. Figures and photos should be original proofs and not xeroxes. Photos must be glossy prints with no screens. Laser prints are not acceptable in place of photos or where gray-scale graphics are involved. All line drawings and photos should be in black and white unless specifically requested. Letters should be large enough to be readily legible when the drawing is reduced to a one-column width as much as 4:1 reduction from the original. Materials should be no larger than 22 x 28 cm (8 1/2 x 11"). Do not incorporate electronic graphics into the paper. List captions on a separate sheet of paper, not on the illustrations.

E. Mandatory Overlength Page Charges

A mandatory page charge is imposed on all papers exceeding eight TRANSACTIONS pages (about 28 double-spaced, typed pages) in length, including illustrations. The charge is \$125 per page for each page over the first eight and is a prerequisite for publication. Details are provided at the time of acceptance; authors are, however, urged to keep this in mind when submitting and revising their papers.

F. Page Charges

After a manuscript has been accepted for publication, the author's company or institution will be approached with a request to pay a charge of \$110 per page to cover part of the cost of publication. Payment of page charges for this IEEE TRANSACTIONS, like journals of other professional societies, is not a necessary prerequisite for publication. The author will receive 100 free reprints (without covers) only if the page charge is honored. Detailed instructions will accompany the proofs.

Scanning the Issue*

Percentile Performance Criteria for Limiting Average Markov Decision Processes, Filar, Krass, and Ross.

This paper addresses the question of whether a policy can be found that achieves a specified value (target) of the long-run limiting average reward at a specified probability level (percentile) for a Markov Decision Process (MDP). This is motivated from the fact that an optimal policy for an infinite horizon MDP with a limiting average reward criterion is insensitive to the probability distribution function of the long-run average reward. Therefore, it is possible that an optimal policy might carry an unacceptably high probability of low values of the long-run average reward. For a problem with a single objective, a complete classification of both the maximal achievable target levels and of their corresponding percentiles is presented. Also, in the case of a communicating MDP, it is shown that every target level can be achieved with only two possible values: zero or one. The approach is constructive and thus furnishes an algorithm for computing a deterministic policy for any feasible target level and percentile pair. Problems with multiple objectives and/or constraints are also studied. These are considered in detail for communicating MDP's with constructive solutions obtained for most cases.

Design and Analysis of Fuzzy Identifiers of Nonlinear Dynamic Systems, Wang.

This paper develops identification algorithms based on "fuzzy" IF-THEN rules for nonlinear dynamic systems. These rules can be either collected from human experts or generated from sensory measurements. Two fuzzy identifiers are developed based upon different rules structures. Both algorithms are shown to be globally stable in the sense that all variables are uniformly bounded and, under mild conditions, the associated prediction errors converge to zero. The algorithms are applied to identify the chaotic glycolytic oscillator.

Stochastic Approximation with Averaging and Feedback: Rapidly Convergent "On Line" Algorithms, Kushner and Yang.

This paper is concerned with the problem of stochastic approximation with averaging of the type considered by Poyak and Juditsky. The authors point out that while the averaging approach has many advantages, choosing the appropriate slowly time-varying gain can

be a nuisance. The authors propose a new stochastic approximation model with constant gain and a feedback term. The rate of convergence of the approximation sequence and the averages are analyzed. This result can be applied to fixed parameter estimation problems. The feedback approach is further extended to time-varying parameter cases. The authors conclude the paper with a simulation study of the proposed algorithms.

Exponential Stabilization of Nonholonomic Chained Systems, Sordalen and Egeland.

This paper analyzes the stabilization of two-input driftless nonholonomic systems in chained form. It is shown how to achieve global asymptotic stability with exponential convergence about any reference configuration by using a nonsmooth time-varying feedback control law. The notion of K-exponential stability is used. Simulation results are given in addition to theoretical analysis.

Global Total Least Squares Modeling of Multivariable Time Series, Roorda and Heij.

This paper concerns the problem of modeling discrete-time multivariate time series by a finite dimensional linear system. The paper differs from other papers on identification in its viewpoint: Model error is defined as the minimal adjustment of the data required to make the time series obey the system equations so that no distinction need be made between system inputs and outputs. The goal of finding an ideal realization can thus be formulated as a total least squares problem. An iterative algorithm is devised to obtain an optimal model, and several numerical simulations are presented as illustrations.

Control of Chained Systems. Application to Path Following and Time-Varying Point Stabilization of Mobile Robots, Samson.

This paper studies the control of a class of nonholonomic kinematic systems. Such systems are nonholonomic in that they are subject to nonintegrable constraints, and they are kinematic in that the controls are velocity controls. The class of systems studied here—systems in chained form—has proven to be particularly amenable to analysis. Here it is shown that systems in chained form may be put into a convenient skew-symmetric chained form, and some open-loop and closed-loop control problems are solved, including global point stabilization by time-varying feedback.

*This section is written by the Transactions Editorial Board.

Percentile Performance Criteria For Limiting Average Markov Decision Processes

Jerzy A. Filar, Dmitry Krass, and Keith W. Ross, *Senior Member, IEEE*

Abstract—In this paper we address the following basic feasibility problem for infinite-horizon Markov decision processes (MDP's): can a policy be found that achieves a specified value (target) of the long-run limiting average reward at a specified probability level (percentile)? Related optimization problems of maximizing the target for a specified percentile and vice versa are also considered. We present a complete (and discrete) classification of both the maximal achievable target levels and of their corresponding percentiles. We also provide an algorithm for computing a deterministic policy corresponding to any feasible target-percentile pair.

Next we consider similar problems for an MDP with multiple rewards and/or constraints. This case presents some difficulties and leads to several open problems. An LP-based formulation provides constructive solutions for most cases.

I. INTRODUCTION AND DEFINITIONS

INFINITE horizon Markov decision processes (MDP's, for short) have been extensively studied since the 1950's. One of the most commonly considered versions is the so-called "limiting average reward" model. In this model the decision-maker aims to maximize the expected value of the limit-average ("long-run average") of an infinite stream of single-stage rewards. There are now a number of good algorithms for computing optimal deterministic policies in the limiting average MDP's (e.g., see [4], [6], [11]).

It should be noted, however, that an optimal policy in the above "classical" sense is insensitive to the probability distribution function of the long-run average reward. That is, it is possible that an optimal policy, while yielding an acceptably high expected long-run average reward, carries with it unacceptably high probability of low values of that same random variable. This "risk insensitivity" is inherent in the formulation of the classical objective criterion as that of maximizing the expected value of a random variable, and it is not necessarily undesirable. Nonetheless, in this paper we adopt the point of view that there are many natural situations where the decision-maker is interested in finding a policy that will achieve a sufficiently high long-run average reward, that is, a target level with a sufficiently high probability, that is, a percentile. The key conceptual difference between this paper

and the classical problem is that our controller is not searching for an optimal policy but rather for a policy that is "good enough," knowing that such a policy will typically fail to exist if the target level and the percentile are set too high. Conceptually, our approach is somewhat analogous to that often adopted by statisticians in testing of hypotheses where it is desirable (but usually not possible!) to simultaneously minimize both the "type 1" and the "type 2" errors. See Bouakiz [5] for a review of similar approaches in economics and operations research literature and White [16] for a review of various approaches to risk-sensitivity in MDP's.

We start out by considering a problem with a single objective. It will be seen (Section IV below) that for our target level-percentile problem it is possible to present a complete (and discrete) classification of both the maximal achievable target levels and of their corresponding percentiles (see Theorem 4.3 and its corollaries). The case of a communicating MDP is particularly interesting as here every target level can be achieved with only two possible values: zero or one (see Theorem 4.1 and its corollary). In all cases our approach is constructive in the sense that we can supply an algorithm for computing a deterministic policy for any feasible target level and percentile pair.

In Section V we turn our attention to problems with multiple objectives and/or constraints. The connection of this to the problem with sample path constraint of [13] and [14] is discussed. In Section VI, we show how the techniques developed in Section IV extend directly to these problems, provided that the problems can be solved for a communicating MDP (see Theorem 6.3). The multiobjective/constrained problems are considered in detail for communicating MDP's, with constructive solutions obtained for most cases (see Theorems 6.1 and 6.2). Conclusions and some open problems are presented in Section VII.

Our analysis is made possible by the recently developed decomposition and sample path theory due to Ross and Varadarajan [14]. The logical development of the results is along the lines of Filar [8]. The latter paper, to the best of our knowledge, introduced the percentile objective criterion in the context of a limiting average Markov control problem, but substituted the long-run expected frequencies in place of actual percentile probabilities since the decomposition and sample path theory of [14] was not known at that time (the unusual way of evaluating risk in [8] was also pointed out in [16]). Some earlier related work appeared in Mitten [12], Sobel [15], and Henig [10]. In the remainder of this section we shall introduce the notation of the limiting average Markov decision process.

Manuscript received January 22, 1992; revised April 12, 1993. This work was supported in part by the AFOSR and the NSF Grants ECS-8704954 and NCR-8707620.

J. A. Filar is with the Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21228 USA.

D. Krass is with the Faculty of Management, University of Toronto, Toronto, Ontario, M5S1V4.

K. W. Ross is with the Department of Systems, University of Pennsylvania, Philadelphia, PA 19104 USA.

IEEE Log Number 9406094.

A finite MDP, Γ , is observed at discrete-time points $n = 1, 2, \dots$. The state space is denoted by $S = \{1, 2, \dots, |S|\}$. With each state $i \in S$ we associate a finite action set $A(i)$. At any time point n , the system is in one of the states, and an action has to be chosen by the decision-maker. If the system is in state i and the action $a \in A(i)$ is chosen, then an immediate reward $r(i, a)$ is earned and the process moves to a state $j \in S$ with transition probability p_{iaj} , where $p_{iaj} \geq 0$ and $\sum_{j \in S} p_{iaj} = 1$.

A decision rule u^n at time n is a function which assigns a probability to the event that action a is taken at time n . In general u^n may depend on all realized states up to and including time n . A policy (or a control) u is a sequence of decision rules: $u = (u^1, u^2, \dots, u^n, \dots)$. A policy is stationary if each u^n depends only on the current state at time n , and $u^1 = u^2 = \dots = u^n = \dots$. A pure (or deterministic) policy is a stationary policy with nonrandomized decision rules. A stationary policy f induces a Markov chain $P(f)$ with the transitions $P(f)_{ij} = \sum_{a \in A(i)} p_{iaj} f_{ia}$ for $i, j \in S$, where f_{ia} is the probability (under f) that action a is chosen whenever state i is visited. If $P(f)$ consists of a single recurrent class of states, then f is called an irreducible policy. If $P(f)$ contains some transient states in addition to a single recurrent class, then f is called a unichain policy. MDP Γ is called unichain if every stationary policy in Γ is unichain.

Let X_n and A_n be the random variables that denote the state at time n and the action chosen at time n , and define the limiting average reward as the random variable

$$R := \lim_{N \rightarrow \infty} \inf \frac{1}{N} \sum_{n=1}^N r(X_n, A_n).$$

It should now be clear that once a policy u and an initial state $X_1 = s_1$ are fixed, the expectation $\phi(u, s_1) := E_u[R | X_1 = s_1]$ of R is well defined and will, from now on, be referred to as the expected average reward due to a policy u . The classical limiting average reward problem is to find an optimal policy u^* such that for all policies u

$$\phi(u^*, s_1) \geq \phi(u, s_1) \quad \text{for all } s_1 \in S. \quad (1.1)$$

It is well known (e.g., see [4]) that there always exists a pure optimal policy u^* .

II. PROBLEMS RELATING TO PERCENTILE OBJECTIVE CRITERIA

We shall say that any pair (τ, α) such that $\tau \in \mathbb{R}$ and $\alpha \in [0, 1]$ constitutes a target level-percentile pair. We shall address the following problems.

Problem 1: Fix $s_1 \in S$. Given $(\tau, \alpha) \in \mathbb{R} \times [0, 1]$ does there exist a policy u such that

$$P_u(R \geq \tau | X_1 = s_1) \geq \alpha? \quad (2.1)$$

If (2.1) holds for some policy u , then we shall say that u achieves the target level τ at percentile α , and τ will be called α -achievable.

Problem 2: Given $\alpha \in [0, 1]$ find

$$\tau_\alpha := \sup \{ \tau \mid \tau \text{ is } \alpha\text{-achievable} \}. \quad (2.2)$$

Problem 3: Given $\tau \in \mathbb{R}$ find

$$\alpha_\tau := \sup \{ \alpha \in [0, 1] \mid \exists \text{ a policy } u \text{ s.t. (2.1) holds} \}. \quad (2.3)$$

Remark 2.1: It should be clear that in many situations the natural goal of maximizing the target level will be in direct conflict with the goal of maximizing the percentile value. This is because τ_α is a nonincreasing function of α , while α_τ is a nonincreasing function of τ .

III. PRELIMINARIES

We shall develop our results within the framework of the decomposition and sample path theory due to Ross and Varadarajan [14] (for a related decomposition, see [2]). This decomposition approach has also proved instrumental in solving other limiting average MDP problems with nonstandard criteria (see [14] and [3]). In this section we collect some results from [14] that will be needed for the proofs in the subsequent sections.

In [14] it is shown that the state space S has a unique partition C_1, C_2, \dots, C_K, T , whose properties are summarized below.

Theorem 3.1 (Proposition 2 of [14]): For any policy u , we have

$$\sum_{k=1}^K P_u(\Phi_k | X_1 = s_1) = 1$$

where

$$\Phi_k := \{X_n \in C_k \text{ almost always}\}$$

(where "almost always" means that $X_n \notin C_k$ finitely often).

The sets C_1, \dots, C_K and T are referred to as strongly communicating classes and the set of transient states, respectively. For a given strongly communicating class C_k , denote by $\Gamma(k)$ the MDP restricted to C_k . Thus, the state space of $\Gamma(k)$ is C_k and the action space $A_k(i)$, $i \in C_k$, is given by

$$A_k(i) = \{a \in A(i) : p_{iaj} = 0 \quad \forall j \notin C_k\}.$$

From [14] we know that $A_k(i)$ is nonempty for all $i \in C_k$ and that $\Gamma(k)$ is a communicating MDP. Recall that a communicating MDP is such that for any pair of states $i, j \in S$, there is a pure policy under which j is accessible from i . Now consider the following linear program $LP(k)$

$$\begin{aligned} \max \quad & \sum_{i \in C_k} \sum_{a \in A_k(i)} r(i, a) x_{ia} \\ \text{s.t.} \quad & \sum_{i \in C_k} \sum_{a \in A_k(i)} (\delta_{ij} - p_{iaj}) x_{ia} = 0 \quad j \in C_k \\ & \sum_{i \in C_k} \sum_{a \in A_k(i)} x_{ia} = 1 \\ & x_{ia} \geq 0, \quad i \in C_k, \quad a \in A_k(i). \end{aligned}$$

Let v_k denote the optimal objective function value of $LP(k)$. We can now state the following result.

Theorem 3.2 (Lemma 4 [14]): For all policies u , all initial states $s_1 \in S$, and all $k = 1, \dots, K$, we have

$$P_u(R \leq v_k \mid \Phi_k, X_1 = s_1) = 1$$

whenever $P_u(\Phi_k, X_1 = s_1) > 0$.

IV. BASIC RESULTS FOR THE SINGLE REWARD CASE

We shall first solve Problems 1–3 for the case of Γ being a communicating MDP. In this case, there is one strongly communicating class C_1 , and T is empty; thus, $S = C_1$.

Consider then $LP(1)$ and to simplify notation denote $v := v_1$. Also let $\{x_{ia}^*\}$ be an optimal basic feasible solution of $LP(1)$ and g^* be a stationary optimal policy constructed from $\{x_{ia}^*\}$ (e.g., see [11], [9], or [13]). Then g^* satisfies

$$\phi(g^*, s_1) = v, \quad s_1 \in S \quad (4.1)$$

where v is the maximal objective function value in $LP(1)$. Moreover, the Markov chain $P(g^*)$ associated with the policy g^* has at most one recurrent class plus (a perhaps empty) set of transient states.

Theorem 4.1: In a communicating MDP Γ there exists a policy that achieves the target level τ with percentile α if and only if $\tau \leq v$. If $\tau \leq v$, then the pure policy g^* achieves the target level τ with percentile α , for any $\alpha \in [0, 1]$.

Note: This result (at least for the unichain case) seems to be well known in the “folklore” of MDP’s. A proof for the communicating case can be found in Asriev and Rotar’ [1] (who establish this result for a much more general stochastic dynamic control model). Since our proof is quite simple, we present it here for completeness.

Proof: Since g^* gives rise to a Markov chain with one recurrent class, we have

$$P_{g^*}(R = \phi(g^*, s_1) \mid X_1 = s_1) = 1$$

(e.g., see [13, Proposition 1 (iii)]). Combining this with (4.1) gives

$$P_{g^*}(R = v \mid X_1 = s_1) = 1. \quad (4.2)$$

From Theorem 3.2, we have

$$P_u(R \leq v \mid X_1 = s_1) = 1. \quad (4.3)$$

The result then follows from (4.2) and (4.3).

As a direct consequence of Theorem 4.1 we have the following corollary.

Corollary 4.1: In a communicating MDP Γ , $\tau_\alpha = v$ for all $\alpha \in (0, 1]$, moreover

$$\alpha_\tau = \begin{cases} 1 & \text{if } \tau \leq v \\ 0 & \text{if } \tau > v. \end{cases}$$

Problems 1–3 have now been solved for the communicating MDP’s. We return to the general case, where we have strongly communicating classes C_1, \dots, C_K and the set T of transient states. Denote by g_k^* the pure policy of Theorem 4.1 associated with $\Gamma(k)$, the MDP restricted to C_k .

Corollary 4.2: For a fixed $k \in \{1, \dots, K\}$, let g be a pure policy that coincides with g_k^* on C_k and is defined arbitrarily elsewhere. Then

$$P_g(R = v_k \mid \Phi_k, X_1 = s_1) = 1$$

if $P_g(\Phi_k, X_1 = s_1) > 0$.

Proof: Note that the definition of Φ_k implies that C_k must be reached in finite time (P_g – a.s.). Thus, the result follows easily from the proof of Theorem 4.1.

Next, we shall consider a fixed target level τ and associate with it an index set $I_\tau = \{k: 1 \leq k \leq K, v_k \geq \tau\}$, and an auxiliary “0-1 MDP” Γ_τ , whose states, actions, and transition probabilities are the same as Γ , but with rewards defined by

$$r^\tau(i, a) := \begin{cases} 1 & \text{if } i \in C_k \text{ and } k \in I_\tau \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that for an arbitrary policy u , the expected average reward in Γ_τ is given by

$$\begin{aligned} \phi^\tau(u, s_1) &= E_u \left[\lim_{N \rightarrow \infty} \inf \frac{1}{N} \sum_{n=1}^N 1(X_n \in C_k) \mid X_1 = s_1 \right] \\ &= \sum_{k \in I_\tau} P_u(\Phi_k \mid X_1 = s_1) \end{aligned} \quad (4.4)$$

where $1(\cdot)$ is the indicator function. Note that the last equality above follows from Theorem 3.1 and the definition of Φ_k , since P_u -a.s. after some finite time t we must have $X_n \in C_k$ for all $n \geq t$ and some $k \in \{1, \dots, K\}$.

Theorem 4.2: Let g^* be an optimal pure policy in Γ_τ which coincides with g_k^* on C_k for $k \in I_\tau$.¹ There exists a policy u satisfying

$$P_u(R \geq \tau \mid X_1 = s_1) \geq \alpha \quad (4.5)$$

where α is the percentile, if and only if $\phi^\tau(g^*, s_1) \geq \alpha$. Further, if the target τ can be achieved at percentile α , then it can be achieved by the pure policy g^* .

Proof: From Theorem 3.1 we have that for any policy u

$$P_u(R \geq \tau \mid X_1 = s_1) = \sum_{k=1}^K P_u(R \geq \tau \mid \Phi_k, X_1 = s_1) P_u(\Phi_k \mid X_1 = s_1). \quad (4.6)$$

From Theorem 3.2 we have

$$P_u(R \geq \tau \mid \Phi_k, X_1 = s_1) = 0 \quad k \notin I_\tau. \quad (4.7)$$

From Corollary 4.2

$$\begin{aligned} 1 &= P_{g^*}(R \geq \tau \mid \Phi_k, X_1 = s_1) \\ &\geq P_u(R \geq \tau \mid \Phi_k, X_1 = s_1) \quad k \notin I_\tau \end{aligned} \quad (4.8)$$

¹ Note that there is no loss of generality here, because g_k^* ensures that once the process enters C_k it remains there forever. Thus it yields the maximal reward of one for every state $i \in C_k$, $k \in I_\tau$.

where the inequality follows from the optimality of g_k^* for $\Gamma(k)$. Combining (4.6)–(4.8) gives

$$\begin{aligned} P_u(R \geq \tau \mid X_1 = s_1) &\leq P_{g^*}(R \geq \tau \mid X_1 = s_1) \\ &= \sum_{k \in I_\tau} P_{g^*}(\Phi_k \mid X_1 = s_1) \\ &= \phi^\tau(g^*, s_1) \end{aligned}$$

from which the result follows.

It is important to note that Theorem 4.2 provides a constructive answer to Problem 1 of Section II concerning α -achievability of the target level τ . We shall now address the problem of determining τ_α —the maximal achievable percentile for the fixed level τ . Towards this goal we assume without loss of generality that the strongly communicating classes C_1, \dots, C_K are ordered so that

$$v_1 \geq v_2 \geq \dots \geq v_K. \quad (4.9)$$

Recall the definition of the MDT Γ_{v_k} (here v_k is the target level). To simplify the notation, we will refer to Γ_{v_k} as Γ_k and to the corresponding expected average reward as ϕ^k instead of ϕ^{v_k} .

Theorem 4.3: Let g_k^* be an optimal pure policy for MDP Γ_k chosen as in Theorem 4.2. We have for $\alpha \in (0, 1]$ that

$$\tau_\alpha = \tau^* := \max \{v_k \mid \phi^k(g_k^*, s_1) \geq \alpha, \quad k = 1, \dots, K\}. \quad (4.10)$$

Proof: Let l be the largest index that achieves the maximum in (4.10), l is well defined since $\phi^K(g_K^*, s_1) = 1$. Since $\tau^* = v_l$, we have $I_{\tau^*} = \{1, 2, \dots, l\}$. Thus, from Theorem 4.2, we know that

$$P_{g^*}(R \geq \tau^* \mid X_1 = s_1) \geq \alpha. \quad (4.11)$$

Hence, τ^* is α -achievable, implying that $\tau_\alpha \geq \tau^*$. If strict inequality were possible in the preceding statement, then there would exist a $\tau' > \tau^*$ and a policy u such that

$$P_u(R \geq \tau' \mid X_1 = s_1) \geq \alpha. \quad (4.12)$$

Now let

$$m := \max \{k : v_k \geq \tau'\}$$

noting that if $v_k < \tau'$ for all $k = 1, \dots, K$, then the left side of (4.12) equals zero contradicting the hypothesis $\alpha > 0$. By the definition of m we have

$$v_m \geq \tau' > \tau^*. \quad (4.13)$$

Applying Theorem 3.1, Theorem 4.1 and optimality g_m^* to (4.12) yields

$$\begin{aligned} \alpha &\leq \sum_{k=1}^K P_u(R \geq \tau' \mid \Phi_k, X_1 = s_1) P_u(\Phi_k \mid X_1 = s_1) \\ &= \sum_{k=1}^m P_u(R \geq \tau' \mid \Phi_k, X_1 = s_1) P_u(\Phi_k \mid X_1 = s_1) \\ &\leq \sum_{k=1}^m P_u(\Phi_k \mid X_1 = s_1) \\ &= \phi^m(u, s_1) \leq \phi^m(g_m^*, s_1). \end{aligned} \quad (4.14)$$

But, by the definition of τ^* , (4.14) implies $\tau^* \geq v_m$, which contradicts (4.13).

Corollary 4.3: The maximal α -achievable target level, τ_α , is a monotone nonincreasing step-function of α , defined on the interval $(0, 1]$.

Proof: Choose g_k^* as in Theorem 4.3. Let $\alpha_k := \phi^k(g_k^*, s_1)$ for $k = 1, \dots, K$, so that $0 \leq \alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_K = 1$. If we define $\tau_0 := v_1$, then by Theorem 4.3, $\tau_\alpha = \tau_0$ for all $\alpha \in (0, \alpha_1]$. Similarly, $\tau_\alpha = \tau_k$, a constant for all $\alpha \in (\alpha_k, \alpha_{k+1}]$, where $\tau_k \geq \tau_{k+1}$ for each $k = 1, \dots, K-1$.

Corollary 4.4: Choose g_k^* as in Theorem 4.3. The maximum percentile α_τ for a given target level τ , is a monotone nonincreasing step function of τ defined in the interval $[v_K, v_1]$. In particular for $\tau \in (v_{k+1}, v_k]$ we have

$$\alpha_\tau = \phi^k(g_k^*, s_1)$$

for each $k = 1, \dots, K-1$.

Proof: This follows easily from the monotonicity of $\phi^k(g_k^*, s_1)$ in the index k .

Remark 4.1: Corollaries 4.3 and 4.4 demonstrate the strength of the percentile objective criteria. Namely, the decomposition of states into C_1, \dots, C_K and T , and the subsequent computation of policies g_k^* together with “breakpoints” τ_k , and v_k for τ_α and α_τ , respectively, allows for a flexible and practical evaluation of gain-risk trade-offs in an average reward MDP.

In view of Corollaries 4.3 and 4.4 the only “reasonable” choices of α and τ are of the special form (τ, α) with $\tau = \tau_\alpha$ and $\alpha = \alpha_\tau$; these correspond to Pareto-optimal solutions.

The preceding results are summarized in the following algorithm, which (for a fixed initial state s_1) finds all target-level-percentile pairs of the indicated “special form”

- Step 1:** Apply the algorithm of [14] to find the decomposition C_1, \dots, C_K, T .
- Step 2:** For each $k \in \{1, \dots, K\}$, find the value v_k of $\Gamma(k)$. Order the strongly communicating classes so that $v_1 \geq \dots \geq v_K$.
- Step 3:** For each $k \in \{1, \dots, K\}$, form the MDP Γ_k and find the optimal policy g_k^* . Let $\alpha_k = \phi^k(g_k^*, s_1)$.
- Step 4:** Let $J = \{(v_k, \alpha_k) \mid k = 1, \dots, K\}$. If $(\tau_1, \alpha), (\tau_2, \alpha) \in J$ and $\tau_1 > \tau_2$, then eliminate (τ_2, α) from J . Continue until no further eliminations are possible.
- Step 5:** We have constructed the set $J = \{(\tau, \alpha) = (\tau_\alpha, \alpha_\tau) \mid \tau \in \mathbb{R}, \alpha \in (0, 1]\}$.

Note that Step 2 is not hard computationally, since very efficient algorithms are known for communicating MDP's. In Step 3 one should use the aggregated MDP method of [14], where each strongly communicating class is replaced by one state. In addition to computational efficiency, this method will automatically yield a deterministic optimal policy satisfying the conditions of Theorem 4.2. Also, since when k is incremented by one, only one immediate reward changes in the aggregated problem (the rest of the data stay the same), perhaps a parametric solution method can be used (e.g. by using LP algorithms for solving the aggregated problem). Further computational efficiency can be gained by quitting

Step 3 as soon as $\alpha_k = 1$ for some k (since all subsequent α_k 's are automatically equal to one). Note also that steps 1 and 2 need not be repeated for different starting states. Finally, we note that before starting the algorithm, one should check whether MDP Γ is communicating (easily verifiable conditions can be found in [9]). If so, the complete characterization is immediately available from Corollary 4.1.

V. PROBLEMS WITH MULTIPLE REWARDS AND CONSTRAINTS—FORMULATIONS

A natural extension of the percentile objective criteria is to the case where each action carries multiple immediate rewards, i.e., each immediate reward is actually a vector of some fixed length L .

$$r(i, a) \in \mathbb{R}^L \text{ for } i \in S, a \in A(i).$$

The definition of the limiting average reward in Section I leads to a vector R of random variables.

At this point two approaches can be taken. Under the first approach, a separate target level-percentile pair would be specified for each of the L components of $R = (R_1, \dots, R_L)$. This leads to the following multi-objective version of Problem 1 (of Section II).

Problem 4: Fix $s_1 \in S$. Given a pair of vectors $(\tau, \alpha) \in \mathbb{R}^L \times [0, 1]^L$, does there exist a policy u such that

$$P_u(R_l \geq \tau_l \mid X_1 = s_1) \geq \alpha_l \text{ for } l = 1, \dots, L?$$

This problem appears to present serious difficulties and is presently unsolved. Below we present an example showing that stationary policies do not suffice for this case, thus indicating that it is unlikely that a simple extension of the results and methods developed in Section IV will work in this case.

Example 5.1: Consider the following MDP with $L = 2$

$$\begin{aligned} r(1, 1) &= (1, 0), & p_{111} &= 1, & r(1, 2) &= (0, 0), & p_{122} &= 1 \\ r(2, 1) &= (0, 1), & p_{212} &= 1. \end{aligned}$$

Thus action 1 is absorbing in both states, and action 2 in state 1 results in immediate rewards of $(0, 0)$ and transition to state 2 with probability one.

Suppose the initial state is one and that the target level-percentile pairs are $(\frac{1}{2}, \frac{1}{2})$ for each component of the reward vector. Note that for any stationary policy f

$$\begin{aligned} P_f(R_1 = 1 \mid X_1 = 1) &= 1, \\ P_f(R_2 = 0 \mid X_1 = 1) &= 1, \text{ if } f_{11} = 1 \end{aligned}$$

and

$$\begin{aligned} P_f(R_1 = 0 \mid X_1 = 1) &= 1, \\ P_f(R_2 = 1 \mid X_1 = 1) &= 1, \text{ if } f_{11} < 1. \end{aligned}$$

Therefore, no stationary policy suffices.

Let g be the deterministic policy that takes action 1 in each state, and let u^1 be the decision rule that takes actions 1 and 2 with probabilities $1/2$ each in state 1. Define the nonstationary

policy u to be $u = (u^1, g)$ (i.e., u uses the decision rule u^1 at time 1 and follows g thereafter). It is easy to see that

$$P_u\left(R_1 \geq \frac{1}{2} \mid X_1 = 1\right) = \frac{1}{2}, \quad P_u\left(R_2 \geq \frac{1}{2} \mid X_1 = 1\right) = \frac{1}{2}$$

and thus u achieves both the specified target-levels at the corresponding percentiles. In fact the same conclusions apply for any $\tau_1, \tau_2 \in (0, 1)$.

Under the second approach to a problem with multiple rewards, a separate target level is specified for each component of R , and all target levels are required to be achieved simultaneously at a single percentile level α .

Problem 5: Fix $s_1 \in S$. Given a vector $\tau \in \mathbb{R}^L$ and $\alpha \in [0, 1]$, does there exist a policy u such that

$$P_u(R_l \geq \tau_l, l = 1, \dots, L \mid X_1 = s_1) \geq \alpha? \quad (5.1)$$

Intuitively, the difference between the two approaches is that in Problem 4, the requirements are placed on the marginal distributions of the components of R , while in Problem 5 the requirement is placed on the joint distribution of R . Henceforth, we will refer to these two approaches as the "marginal-probability" and the "joint-probability" formulations, respectively. While the marginal-probability formulation provides for more modeling flexibility, the advantage of the joint-probability formulation is that it leads to more tractable problems: most of the results obtained for Problem 1 will be extended to Problem 5.

We note that an extension to multiple rewards is especially important, since some of the components of $r(i, a)$ can be regarded as negatives of the costs. In this case, the corresponding components of R and τ can be regarded as constraints which must be satisfied at the specified percentile level (either singly in Problem 4 or jointly in Problem 5). This can be seen as a generalization of the "sample path constraint" of Ross and Varadarajan [13] and [14] (which in our notation corresponds to $\alpha = 1$ and $L = 1$).

VI. BASIC RESULTS FOR THE MULTIPLE REWARDS CASE

In this section we consider Problem 5 for the case of communicating MDP's. We employ the linear-programming-based techniques developed in [6], [11], and [13] and introduced in Section III above. It should be noted that the problem considered in this section is related to the problem of finding optimal policies in average-reward communicating MDP's with state-action frequency constraints. It is well known (see e.g., [7, example 9.3]) that stationary policies may not be optimal in this case. For a problem with one inequality constraint, [13] shows that an ϵ -optimal stationary policy can always be found (the policy is generally not deterministic). This result is related to Theorem 6.2.

In the sequel, all vector inequalities and equalities are defined componentwise, that is $R \geq \tau$ means $R_l \geq \tau_l$, $l = 1, \dots, L$.

We first state some preliminary definitions and results. Consider the following polyhedral set

$$\Delta = \left\{ \begin{array}{l} \sum_{a \in A(i)} x_{ia} = \sum_{j \in S} \sum_{a \in A(j)} p_{ja} x_{ja}, \quad i \in S \\ \sum_{i \in S} \sum_{a \in A(i)} x_{ia} = 1 \\ x_{ia} \geq 0, \quad i \in S, a \in A(i) \end{array} \right. \quad (6.1)$$

(this is simply the feasible region of $LP(k)$ of Section III). For $\mathbf{x} = (x_{ia}) \in \Delta$, define a stationary policy

$$f(\mathbf{x})_{ia} = \begin{cases} \frac{x_{ia}}{\sum_{a \in A(i)} x_{ia}} & \text{if } i \in S, \sum_{a \in A(i)} x_{ia} > 0, \\ & a \in A(i) \\ \frac{1}{|A(i)|} & \text{if } i \in S, \sum_{a \in A(i)} x_{ia} = 0. \end{cases} \quad (6.2)$$

Define the following random variable (whenever it exists), representing long-term average state-action frequencies

$$Z_{ia} = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T 1[X_t = i, A_t = a]. \quad (6.3)$$

We will need the following well-known results, the proofs of which can be found in [11] and [13].

Lemma 6.1: Let Γ be an arbitrary MDP.

i) If f is a stationary policy then P_f -almost surely, Z is well-defined, $Z \in \Delta$, and

$$R = \sum_{i \in S} \sum_{a \in A(i)} r(i, a) Z_{ia}.$$

ii) If f is a stationary policy and f is unichain, then

$$Z_{ia} = \pi(f)_i f_{ia} \quad P_f - \text{a.s.}$$

where $\pi(f)$ is the unique stationary probability vector of $P(f)$.

iii) If $\mathbf{x} \in \Delta$ and $f(\mathbf{x})$ is unichain, then

$$Z = \mathbf{x} P_{f(\mathbf{x})} - \text{a.s.}$$

iv) If Γ is a communicating MDP and $\mathbf{x} \in \Delta$, $\mathbf{x} > 0$, then $f(\mathbf{x})$ is an irreducible stationary policy.

Let τ be a given target vector and α a given percentile level.

We now define the following linear program LP

max b

Subject to

$$\begin{array}{l} \sum_{a \in A(i)} x_{ia} = \sum_{j \in S} \sum_{a \in A(j)} p_{ja} x_{ja} \quad i \in S \\ \sum_{i \in S} \sum_{a \in A(i)} x_{ia} = 1 \\ \sum_{i \in S} \sum_{a \in A(i)} r(i, a) x_{ia} \geq \tau_l \quad l = 1, \dots, L \\ x_{ia} \geq b, \quad i \in S, a \in A(i) \\ b \geq 0. \end{array}$$

Note that the feasible region of LP is contained in $\Delta \times \mathbf{R}$.

Theorem 6.1: Let Γ be a communicating MDP.

i) If LP is infeasible, then for any policy u

$$P_u(R \geq \tau \mid X_1 = s_1) = 0.$$

ii) Suppose LP is feasible. Let (\mathbf{x}^*, b^*) be an optimal solution and $f^* = f(\mathbf{x}^*)$. If $b^* > 0$, or f^* is unichain, then

$$P_{f^*}(R \geq \tau \mid X_1 = s_1) = 1.$$

Proof:

- i) The proof is analogous to the proof of Proposition 2 in [14].
- ii) If $b^* > 0$ then f^* is irreducible by Lemma 6.1-iv) and thus, by Lemma 6.1-iii), $\mathbf{x}^* = Z P_{f^*}$ -a.s. It now follows by Lemma 6.1-i) and the feasibility of \mathbf{x}^* that

$$R \geq \tau P_{f^*} - \text{a.s.}$$

This result is a counterpart of Theorem 4.1 for the single-objective case. When conditions in part ii) of Theorem 6.1 hold, it provides a solution to Problem 5 for any $\alpha \in [0, 1]$, and if the condition in part i) holds then Problem 5 has no solutions for any α . Unlike Theorem 4.1, however, the current result does not provide a complete characterization of solutions. If the optimal value b^* of LP is equal to zero and $f(\mathbf{x}^*)$ is not unichain, then it is not currently known whether Problem 5 has any solutions or any solutions in stationary policies. The following example shows that it is possible to have a situation where no feasible stationary policies exist when $b^* = 0$.

Example 6.1: Consider the following MDP Γ with $L = 2$

$$\begin{array}{cc} i = 1 & i = 2 \\ \begin{array}{l} r(1, 1) = (1, 0), \quad p_{111} = 1 \\ r(1, 2) = (0, 0), \quad p_{122} = 1 \end{array} & \begin{array}{l} r(2, 1) = (0, 1), \quad p_{212} = 1 \\ r(2, 2) = (0, 0), \quad p_{221} = 1 \end{array} \end{array}$$

where $s_1 = 1$, $\tau = (1/2, 1/2)$ and $\alpha > 0$. Then LP is feasible with $b^* = 0$ and

$$\mathbf{x}^* = (x_{11}, x_{12}, x_{21}, x_{22})^T = (1/2, 0, 1/2, 0)^T.$$

Thus $f(\mathbf{x}^*)$ is not unichain, and in fact it is easy to see that for any stationary policy f , $P_f(R \geq \tau) = 0$. It is not known whether there exists some nonstationary policy u such that $P_u(R \geq \tau) > 0$.

It might be reasonable to suppose that when $b^* = 0$, the presence of feasible stationary policies might be detected by checking whether $f(\mathbf{x}^*)$ is unichain for any optimal basic feasible solution \mathbf{x}^* of LP (this would cover Example 6.1 where there is only one optimal basic feasible solution). The following example, however, shows that it is possible that $b^* = 0$, $f(\mathbf{x}^*)$ is not unichain for any optimal basic feasible solution, and yet, feasible stationary policies exist for any $\alpha > 0$.

Example 6.2: Consider the following MDP Γ with $L = 3$

$$\begin{array}{cc} i = 1 & i = 2 \\ \begin{array}{l} r(1, 1) = (1, 0, 0), \quad p_{111} = 1 \\ r(1, 2) = (0, 0, 0), \quad p_{122} = 1 \\ r(1, 3) = (-1, 0, 0), \quad p_{131} = 1 \end{array} & \begin{array}{l} r(2, 1) = (0, 1, 0), \quad p_{211} = 1 \\ r(2, 2) = (0, 1, 0), \quad p_{223} = 1 \\ r(2, 3) = (0, 1, 0), \quad p_{232} = 1 \end{array} \\ i = 3 & \\ \begin{array}{l} r(3, 1) = (0, 0, 1), \quad p_{313} = 1 \\ r(3, 2) = (0, 0, 0), \quad p_{322} = 1 \end{array} & \end{array}$$

and let $s_1 = 1$. Take some $\alpha > 0$ and $\tau = (1/4, 1/4, 1/4)$. It is not hard to verify that the resulting LP has the optimal value $b^* = 0$ with only two optimal basic feasible solutions

$$\begin{aligned} \mathbf{x}^1 &= (x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{31}, x_{32})^T \\ &= (1/4, 1/4, 0, 1/4, 0, 1/4, 0)^T \end{aligned}$$

and

$$\mathbf{x}^2 = (1/4, 0, 0, 0, 1/4, 1/4, 1/4)^T.$$

Neither $f(\mathbf{x}^1)$ nor $f(\mathbf{x}^2)$ is unichain and, in fact, the probability of achieving τ is zero for both of these policies. If we take the feasible point

$$\mathbf{x}^* = \frac{1}{2}\mathbf{x}^1 + \frac{1}{2}\mathbf{x}^2 = (1/4, 1/8, 0, 1/8, 1/8, 1/4, 1/8)^T$$

then $f(\mathbf{x}^*)$ is unichain and $P_f(\mathbf{x}^*)(R \geq \tau \mid X_1 = s_1) = 1$

We will call a target vector τ for which LP is feasible and has optimal value $b^* = 0$ an indeterminate target vector.

We note that this case does not arise in the case of unichain MDP's (since one of the two conditions of Theorem 6.1 must be met), and we have the following corollary.

Corollary 6.1: Suppose Γ is a unichain MDP. Then either LP is infeasible, in which case Problem 5 has no solutions for any $\alpha > 0$, or LP is feasible with an optimal solution (b^*, \mathbf{x}^*) , in which case for $f = f(\mathbf{x}^*)$, $P_f(R \geq \tau \mid X_1 = s_1) = 1$.

In the remainder of this section we show that the difficulties associated with indeterminate target vectors can always be avoided by a slight relaxation of the target vector and that indeterminate target vectors are sufficiently "rare" in the set of all possible target vectors. We will need the following result.

Lemma 6.2: Suppose Γ is a communicating MDP. Take $\mathbf{x} \in \Delta$. Then for any $\epsilon > 0$ and $s_1 \in S$, there exists an irreducible stationary policy f such that

$$P_f(\|\mathbf{Z} - \mathbf{x}\| \leq \epsilon \mid X_1 = s_1) = 1$$

(where $\|\cdot\|$ is an arbitrary vector norm).

Proof: If $\mathbf{x} > 0$, then $f = f(\mathbf{x})$ is irreducible by Lemma 6.1-iv), and it follows by Lemma 6.1-iii) that $\mathbf{Z} = \mathbf{x}$ P_f -a.s.

Now suppose that $x_{ia} = 0$ for some $i \in S$, $a \in A(i)$. Let g be a completely randomized (stationary) policy (i.e., $g_{ia} = \frac{1}{|A(i)|}$ for any $i \in S$, $a \in A(i)$). Since Γ is a communicating MDP, g must be irreducible. By Lemma 6.1-ii), $\mathbf{Z} = \mathbf{x}(g)$ P_g -almost surely, where $\mathbf{x}(g)_{ia} = \pi(g)_{ia}$ for $i \in S$, $a \in A(i)$. By Lemma 6.1-i), $\mathbf{x}(g) > 0$ and $\mathbf{x}(g) \in \Delta$. Let

$$\mathbf{x}(\lambda) = \lambda \mathbf{x} + (1 - \lambda)\mathbf{x}(g) \text{ for } \lambda \in (0, 1).$$

Clearly, $\mathbf{x}(\lambda) \in \Delta$ and $\mathbf{x}(\lambda) > 0$ for any $\lambda < 1$. It follows by continuity of $\mathbf{x}(\lambda)$ with respect to λ that we can choose $\lambda^* \in (0, 1)$ so that

$$\|\mathbf{x}(\lambda^*) - \mathbf{x}\| \leq \epsilon. \quad (6.4)$$

Let $f = f(\mathbf{x}(\lambda^*))$. By Lemma 6.1-iv), f must be irreducible. It follows by Lemma 6.1-iii) that $\mathbf{Z} = \mathbf{x}(\lambda^*)$ P_f -a.s. The result now follows immediately from (6.4).

We are now ready to prove the following result.

Theorem 6.2: Suppose Γ is a communicating MDP and τ is an indeterminate target vector. Let δ be an arbitrary vector with positive components and choose $\epsilon > 0$. Then there exists

an irreducible stationary policy f such that

$$P_f(R \geq \tau - \epsilon \delta \mid X_1 = s_1) = 1.$$

Proof: Let (b^*, \mathbf{x}^*) be an optimal solution of LP with target vector τ . By assumption, $b^* = 0$. By Lemma 6.2, for any $\gamma > 0$, there exists an irreducible stationary policy f such that $\|\mathbf{x}' - \mathbf{x}^*\| \leq \gamma$, where $\mathbf{x}' = \mathbf{Z}$ is a constant P_f -almost surely. Choose γ small enough so that

$$\sum_{i \in S} \sum_{a \in A(i)} x'_{ia} \tau(i, a) \geq \tau - \epsilon \delta \quad P_f \text{ a.s.} \quad (6.5)$$

(this can always be done since the left-hand side of (6.5) is a continuous function of \mathbf{x}' and $\sum_{i \in S} \sum_{a \in A(i)} x'_{ia} \tau(i, a) \geq \delta$).

Since f is irreducible, by Lemma 6.1-ii), P_f -almost surely $\mathbf{x}' > 0$, i.e., $\mathbf{x}' \geq b' > 0$ for some $b' \in \mathbb{R}$. It follows from (6.5) that (b', \mathbf{x}') is feasible in LP with the target vector $\tau - \epsilon \delta$ and consequently, the optimal value for this LP is positive. The result now follows from Theorem 6.1-ii).

Thus, any strict relaxation of an indeterminate target vector produces a target vector for which Problem 5 can be solved by an irreducible policy for any percentile level α . Geometrically, the situation is as follows: let $T = \{\tau \mid LP \text{ is feasible}\}$. T must be a closed set. Let

$TS = \{\tau \mid \text{Problem 5 has a solution in unichain policies}$

for any percentile $\alpha \in [0, 1]\}$,

$TF = \{\tau \mid LP \text{ has optimal value } b^* > 0\}$.

$TI = \{\tau \mid \tau \text{ is an indeterminate target vector}\}$.

Then $TS \cup TI = T$, $TF \subset TS$, and $TI \subset \text{boundary}\{T\}$. Note also that the intersection of TS and TI need not be empty, as shown by Example 6.2. We illustrate the relationships between these sets in the context of Example 6.1, where

$$T = \{\tau_1 \leq 1, \tau_2 \leq 0\} \cup \{\tau_1 \leq 0, \tau_2 \leq 1\} \\ \cup \{\tau_1 + \tau_2 \leq 1, \tau_1 > 0, \tau_2 > 0\}$$

(represented by the shaded region on Fig. 1)

$$TS = T \setminus \{\tau_1 + \tau_2 = 1, \tau_1 > 0, \tau_2 > 0\},$$

$$TI = \{\tau_1 = 1, \tau_2 \leq 0\} \cup \{\tau_1 \leq 0, \tau_2 = 1\} \\ \cup \{\tau_1 + \tau_2 = 1, \tau_1 > 0, \tau_2 > 0\}$$

(the boundary of T), and

$$TF = T \setminus TI.$$

Remark 6.1: In summary, our results for the communicating case with multiple rewards closely parallel the corresponding results for the single reward: for "most" target vectors, either Problem 5 has no solution or else it is solvable by irreducible policies for any percentile level. The differences from the single reward case lie in the fact that the feasible stationary policies might have to be randomized (for single reward deterministic policies sufficed), and in our inability to

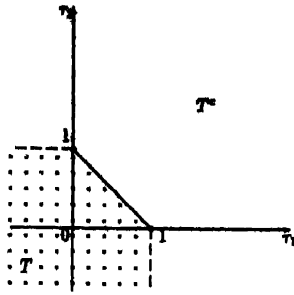


Fig. 1. The set of possible target vectors for Example 6.1.

handle (completely) the indeterminate target vector case. If the modeler has some flexibility in setting the target vector, the indeterminate case can always be avoided. In some cases, however, such flexibility might not exist. Therefore, we consider the further study of the indeterminate case to be important. Specifically, the following questions should be addressed:

- 1) Must nonstationary feasible policies exist when τ is indeterminate? In particular, do such policies exist in Example 6.1?
- 2) Can the indeterminate target vectors for which Problem 5 has a solution in unichain policies (i.e., $\tau \in TI \cap TS$) be characterized? It would be particularly interesting to find a computationally simple characterization or show that one does not exist.

We now turn our attention to the general (multichain) MDPs. Surprisingly, the extension of our communicating MDP results to this case can be done quite easily by employing the decomposition and sample path theory.

As in Section IV above, let C_1, \dots, C_K, T be the strongly communicating classes and the set of transient states of Γ , and let $\Gamma(k)$ be the MDP restricted to C_k . Define the index set

$$J_\tau = \{k \mid 1 \leq k \leq K, \text{ The optimal value of } LP \text{ for } \Gamma(k) \text{ is positive}\}.$$

For $k \in J_\tau$, let f_k^* be the stationary policy of Theorem 6.1-ii) associated with the MDP $\Gamma(k)$. Following Section IV, define MDP Γ_τ whose states, actions, and the transition law are the same as Γ , but with the rewards defined by

$$r^\tau(i, a) := \begin{cases} 1 & \text{if } i \in C_k, a \in A(i) \text{ and } k \in J_\tau \\ 0 & \text{otherwise.} \end{cases} \quad (6.6)$$

Theorem 6.3: Assume the target vector τ is not an indeterminate target vector for any $\Gamma(k)$, $k = \{1, \dots, K\}$. Let f^* be an optimal stationary policy in Γ_τ which coincides with f_k^* on C_k for $k \in J_\tau$.² There exists a policy u satisfying

$$P_u(R \geq \tau \mid X_1 = s_1) \geq \alpha \quad (6.7)$$

if and only if $\phi^\tau(f^*, s_1) \geq \alpha$. Further, if the target vector τ can be achieved at percentile α , then it can be achieved by the stationary policy f^* .

Proof: Exactly the same as in Theorem 4.2.

This result provides a constructive answer to Problem 5, provided τ satisfies the assumption of Theorem 6.3.

²See the footnote in Theorem 4.2.

Remark 6.2: Similarly to Problem 3 of Section II, define

$$\alpha_\tau = \sup \{\alpha \in [0, 1] \mid \exists \text{ a policy } u \text{ s.t. (5.1) holds}\}.$$

It is clear from the proof of Theorem 4.2 that $\alpha_\tau = \phi^\tau(f^*, s_1)$ and the policy f^* constructed above achieves τ at percentile α_τ .

VII. PROBLEMS FOR FURTHER RESEARCH

In the preceding sections we have outlined several problems for further research. For the multiple reward case, a satisfactory treatment of the marginal-probability formulation described in Section V would be very useful. Open problems connected with the indeterminate target vector case were discussed in Remark 6.1.


Another important open problem associated with the percentile objective criterion is the satisfactory treatment of the discounted case. Analysis of a similar problem for this case can be found in Bouakiz [5]; however, no algorithmic results were obtained. Problems 1–3 of Section II appear to be much harder in this case, mainly because no equivalent of the decomposition theory is known (or perhaps, possible) for discounted MDP's.

ACKNOWLEDGMENT

The authors thank M. Teboulle for his helpful discussions.

REFERENCES


- [1] A. V. Asriev and V. I. Rotar', "On asymptotic optimality in probability and almost surely in dynamic control," *Stochastics and Stochastics Rep.*, vol. 33, pp. 1–16, 1990.
- [2] J. Bather, "Optimal decision procedures in finite Markov chains. Part III: General convex systems," *Advances in Applied Prob.*, vol. 5, pp. 541–553, 1973.
- [3] M. Bayal-Gursoy and K. W. Ross, "Variability sensitive Markov decision processes," *Mathematics Oper. Res.*, vol. 17, no. 3, pp. 558–572, 1992.
- [4] D. Blackwell, "Discrete dynamic programming," *Annals Math. Stat.*, vol. 33, pp. 719–726, 1962.
- [5] M. Bouakiz, "Risk sensitivity in stochastic optimization with applications," Ph.D. dissertation, Georgia Inst. Tech., Atlanta, 1985.
- [6] C. Derman, *Finite State Markovian Decision Processes*. New York: Academic, 1970.
- [7] E. A. Feinberg, "Constrained semi-Markov decision processes with average rewards," working paper, State Univ. of New York, Stony Brook, 1992.
- [8] J. A. Filar, "Percentiles and Markovian decision processes," *Oper. Res. Lett.*, vol. 2, pp. 13–15, 1983.
- [9] J. A. Filar and T. A. Schulz, "Communicating MDPs: equivalence and LP properties," *Op. Res. Lett.*, vol. 7, pp. 303–307, 1988.
- [10] M. I. Henig, "The principle of optimality in dynamic programming with returns in partially ordered sets," *Math. Op. Res.*, vol. 10, pp. 462–470, 1985.
- [11] L. C. M. Kallenberg, *Linear Programming and Finite Markovian Control Problems*. Amsterdam: Mathematical Center Tracts 148, 1983.
- [12] L. G. Mitten, "Preference order dynamic programming," *Management Science*, vol. 21, pp. 43–46, 1975.
- [13] K. W. Ross and R. Varadarajan, "Markov decision processes with sample path constraints: The communicating case," *Op. Res.*, vol. 37, pp. 780–790, 1989.
- [14] ———, "Multichain Markov decision processes with a sample path constraint: A decomposition approach," *Math Op. Res.*, vol. 16, no. 1, pp. 195–207, 1991.
- [15] M. J. Sobel, "Ordinal dynamic programming," *Management Science*, vol. 21, pp. 967–975, 1975.
- [16] D. J. White, "Mean, variance, and probabilistic criteria in finite Markov decision processes: A review," *JOTA*, vol. 56, no. 1, pp. 1–29, 1988.



Jerzy A. Filar was born in Warsaw in 1949. He studied Mathematics and Statistics at the University of Melbourne, Australia and at Monash University. He received the Ph.D. degree at the University of Illinois, Chicago, in 1980.

Dr. Filar is currently Professor of Mathematics and Statistics at the University of South Australia and is Director of its Centre for Mathematical Applications. He has also held academic positions at the University of Minnesota, the Johns Hopkins University, and the University of Maryland, Baltimore County.

His current research interests include operations research, control theory, and environmental modeling.



Dmitry Krass received the B.Sc. in mathematics from the University of Chicago in 1984 and the M.S.E. and the Ph.D. in operations research from Johns Hopkins University in 1986 and 1989, respectively.

Since 1989, Dr. Krass has been as Associate Professor of Operations Management and Statistics at the Faculty of Management, University of Toronto. His current research interests include optimization and stochastic dynamic programming, and in developing advanced tools for managerial decision

support in stochastic environments.

Kelth W. Ross (S'82-M'85-SM'90) received the B.S. degree from Tufts University, Medford, MA, in 1979, the M.S. degree from Columbia University, New York, in 1981, and the Ph.D. degree from the University of Michigan, Ann Arbor, in 1985.

He is an Associate Professor in the Department of Systems Engineering, University of Pennsylvania, Philadelphia. He also holds secondary appointments in the Computer Information Science and in the Operations and Informations Management (Wharton) Departments. In 1980 he designed satellite radar systems as an employee of AVCO. He has been a visiting scholar at several research and academic institutions in France. His current research interests include protocols and traffic management in high-speed telecommunication networks, including local area networks, wide area data networks, voice networks, and broadband integrated services digital networks.

Dr. Ross is the Program Chairman of the 1995 ORSA Telecommunications Conference and is an Associate Editor for *Probability in the Engineering and the Information Sciences* and for *Telecommunication Systems*. He has published over 30 papers in leading journals and is completing a book on multiservice loss models for broadband telecommunication networks. He is the recipient of numerous grants from AT&T and NSF.

Design and Analysis of Fuzzy Identifiers of Nonlinear Dynamic Systems

Li-Xin Wang, *Member, IEEE*

Abstract—In this paper, we use fuzzy systems as identifiers for nonlinear dynamic systems. We provide a theoretical justification for the fuzzy identifiers by proving that they are capable of following the output of a general nonlinear dynamic system to arbitrary accuracy in any finite time interval. The fuzzy identifiers are constructed from a set of adaptable fuzzy IF-THEN rules and can combine both numerical information (in the form of input-output pairs obtained by exciting the system with an input signal and measuring the corresponding outputs) and linguistic information (in the form of IF-THEN rules about the behavior of the system in terms of vague and fuzzy words) into their designs in a uniform fashion. We develop two fuzzy identifiers. The first one is designed through the following four steps: 1) define some fuzzy sets in the state space $U \subset R^n$ of the system; these fuzzy sets do not change; 2) construct fuzzy rule bases of the fuzzy identifier which comprise rules whose IF parts constitute all the possible combinations of the fuzzy sets defined in 1); 3) design the fuzzy systems in the fuzzy identifier based on the fuzzy rule bases of 2); and 4) develop an adaptive law for the free parameters in the fuzzy identifier. The second fuzzy identifier is designed in a similar way as the first one except that: a) the parameters characterizing the fuzzy sets in the state space change during the adaptation procedure; and b) the fuzzy systems and the adaptive law are different. We prove that: 1) both fuzzy identifiers are globally stable in the sense that all variables in the fuzzy identifiers are uniformly bounded, and 2) under some conditions the identification errors of both fuzzy identifiers converge to zero asymptotically. Finally, we simulate the fuzzy identifiers for identifying the chaotic glycolytic oscillator, and the results show that: 1) the fuzzy identifiers can approximate the chaotic system at a reasonable speed and accuracy without using any linguistic information, and 2) by incorporating some fuzzy linguistic IF-THEN rules about the behavior of the system into the fuzzy identifiers, the speed and accuracy of the fuzzy identifiers are greatly improved.

I. INTRODUCTION

EXISTING identification schemes determine a model for a system based on the input-output pairs which are collected by exciting the system with an input signal and measuring the corresponding outputs [8], [20]. For many complex systems, however, an important portion of information comes from another source: human experts, who are very familiar with the systems and can provide linguistic descriptions about the behavior of the system in terms of vague and fuzzy words. For example, although we may not know the exact mathematical model of a car, we can describe the behavior of

the car as follow: "IF apply 'more' force to the accelerator of the car, THEN the speed of the car will 'increase'," where the 'more' and 'increase' are characterized by fuzzy membership functions. Although these linguistic descriptions are not precise, they provide important information about the system. In fact, for many industrial process control problems, a human operator can determine a set of successful control rules based only on the linguistic descriptions about the processes [5], [11], [13], [17], [38]. Unfortunately, existing identification schemes ignore this important source of information and cannot incorporate the linguistic descriptions directly into the identifiers. The goal of this paper is to develop identifiers of nonlinear systems which combine both linguistic descriptions and input-output pairs in a uniform fashion into their designs.

Fuzzy systems are powerful tools for achieving this goal because, on one hand, fuzzy systems are constructed from fuzzy IF-THEN rules so that linguistic descriptions can be naturally incorporated into the fuzzy identifiers; on the other hand, adaptive laws can be developed (as shown in this paper) to adjust the free parameters of the fuzzy identifiers to make them match the input-output pairs. Specifically, the initial identifiers are constructed from linguistic information, and the parameters are then adjusted on-line based on input-output pairs. In this way, the fuzzy identifiers combine both linguistic and numerical information into their designs.

The idea of fuzzy systems was introduced by Zadeh [41] very early in the literature of fuzzy sets. Research on fuzzy systems has been developed in two main directions. The first studies fuzzy systems in the same conceptual framework as classical systems and has given birth to a body of abstract results concerning the stability [4], [14], reachability [6], [25], observability [6], [25], etc., of fuzzy systems. The second direction is the linguistic approach to fuzzy systems [22], [30], [43], [44], in which a fuzzy system is constructed from a collection of fuzzy rules (propositions) and a fuzzy inference engine which uses techniques in approximate reasoning [1], [45]. This approach was used to synthesize fuzzy logic controllers [18], [19], [22] which has been successfully applied to a wide variety of practical problems during the past decade [5], [11], [13], [17], [30], [40]. We adapt this second approach to fuzzy systems in this paper.

The basic configuration of a fuzzy system is shown in Fig. 1. There are four components in a fuzzy system:

1) Fuzzy rule base which comprises fuzzy rules describing how the fuzzy system performs; it is the heart of the whole system in the sense that other three components are used to interpret and combine these rules to form the final system.

Manuscript received February 6, 1992; revised January 11, 1994 and April 30, 1994. Recommended by Associate Editor, S. P. Meyn.

The author is with the Department of Electrical and Electronic Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

IEEE Log Number 9406095.

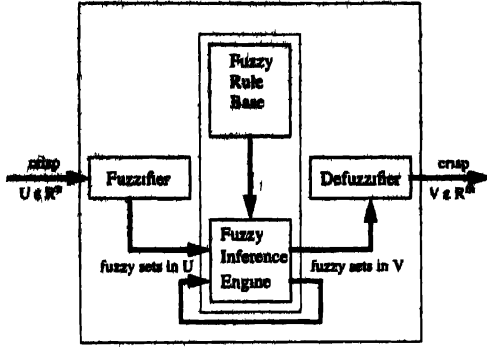


Fig. 1. Basic configuration of fuzzy systems.

2) Fuzzy inference engine which uses techniques in approximate reasoning to determine a mapping from the fuzzy sets in the input space $U \subset R^n$ to the fuzzy sets in the output space $V \subset R^m$.

3) Fuzzifier which maps crisp points in the input space into fuzzy sets in the input space.

4) Defuzzifier which maps fuzzy sets in the output space into crisp points in the output space. Depending upon whether there are fuzzifier and defuzzifier, we have two classes of fuzzy systems. The first class of fuzzy systems comprise only the fuzzy rule base and fuzzy inference engine and therefore operate in a pure fuzzy environment, i.e., inputs and outputs to these fuzzy systems are fuzzy variables. If there is a feedback (as shown in Fig. 1 by the dashed arrow line), we have the so-called fuzzy dynamic systems [4], [14], [32], [33]. The second class of fuzzy systems comprise all the four components and perform mappings from crisp $U \subset R^n$ to crisp $V \subset R^m$. In the literature, these second-class fuzzy systems are called fuzzy logic controllers [18], [19], because their most successful applications have been to control problems. In this paper, we use these second-class fuzzy systems as identifiers for nonlinear dynamic systems. Also, we consider only static fuzzy systems.

In Section II, we present a detailed description for each component of the fuzzy system and show the universal approximation properties of the fuzzy systems to nonlinear static and dynamic systems. In Sections III and IV, we develop two identifiers of nonlinear systems based on the fuzzy system models and study their stability and convergence properties. In Section V, we simulate the two fuzzy identifiers for a chaotic system and show how identification performance is improved by incorporating linguistic descriptions about the system into the fuzzy identifiers. Section VI concludes the paper.

II. DESCRIPTION AND ANALYSIS OF FUZZY SYSTEMS

Fig. 1 shows the basic configuration of the fuzzy systems considered in this paper. The fuzzy system performs a mapping from crisp $U \subset R^n$ to crisp $V \subset R^m$. We assume that $U = U_1 \times \cdots \times U_n$ and $V = V_1 \times \cdots \times V_m$, where $U_i, V_j \subset R$, $i = 1, 2, \dots, n$, and $j = 1, 2, \dots, m$. We first briefly review some basic concepts in fuzzy sets and systems theory which are useful in describing the fuzzy systems and then present a detailed description for each component of the fuzzy system.

A fuzzy set [42] F of a universe of discourse U is characterized by a membership function $\mu_F: U \rightarrow [0, 1]$ which associates with each element u of U a number $\mu_F(u)$ in the interval $[0, 1]$ which represents the grade of membership of u in F . The label F of a fuzzy set is often some linguistic term like "small," "large," etc.. A linguistic variable [44] is a variable whose values are words in a natural or artificial language, and these words are labels of fuzzy sets. A fuzzy IF-THEN rule [44] is an expression of the form "IF A THEN B ," where A and B are statements about what values the linguistic variables take on. A triangular norm ' \star ' is defined as a mapping from $[0, 1] \times [0, 1]$ to $[0, 1]$, and the most commonly used triangular norms are [18]

$$\begin{aligned} \min[x, y] & \quad \text{intersection} \\ x \star y &= xy \quad \text{algebraic product} \\ \max[0, x + y - 1] & \quad \text{bounded product} \end{aligned} \quad (1)$$

where $x, y \in [0, 1]$. A triangular conorm ' $\dot{+}$ ' is a mapping from $[0, 1] \times [0, 1]$ to $[0, 1]$, and the most commonly used triangular conorms are [18]

$$\begin{aligned} \max[x, y] & \quad \text{union} \\ x + y &= x + y - xy \quad \text{algebraic sum} \\ \min[1, x + y] & \quad \text{bounded sum.} \end{aligned} \quad (2)$$

Let F_1, \dots, F_n be fuzzy sets in X_1, \dots, X_n , respectively; the cartesian product of F_1, \dots, F_n , denoted by $F_1 \times \cdots \times F_n$, is a fuzzy set in the product space $X = X_1 \times \cdots \times X_n$ with membership function

$$\mu_{F_1 \times \cdots \times F_n}(x_1, \dots, x_n) = \mu_{F_1}(x_1) \star \cdots \star \mu_{F_n}(x_n). \quad (3)$$

Let R and S be fuzzy sets in $X \times W$ and $W \times Y$, respectively; the supstar composition of R and S , denoted by $R \circ S$, is a fuzzy set in $X \times Y$ with membership function

$$\mu_{R \circ S}(x, y) = \sup_{w \in W} [\mu_R(x, w) \star \mu_S(w, y)]. \quad (4)$$

Let A and B be fuzzy sets in X and Y , respectively; the fuzzy implication $A \rightarrow B$, or 'IF A THEN B ', is defined as a fuzzy set in $X \times Y$ with membership function

$$\mu_{A \rightarrow B}(x, y) = \mu_A(x) \star \mu_B(y). \quad (5)$$

Operations other than \star may be used in (5) to define the fuzzy implication (see [18]).

A fuzzy rule base R is a collection of fuzzy IF-THEN rules

$$R = [R^1, R^2, \dots, R^M] \quad (6)$$

with

$$\begin{aligned} R^l: & \text{ IF } (x_1 \text{ is } F_1^l \text{ and } \cdots \text{ and } x_p \text{ is } F_p^l); \\ & \text{ THEN } (y_1 \text{ is } G_1^l, \dots, y_q \text{ is } G_q^l) \end{aligned} \quad (7)$$

where $\underline{x} = (x_1, \dots, x_n)^T$ and $\underline{y} = (y_1, \dots, y_m)^T$ are the input and output vectors to the fuzzy system, respectively, F_i^l and G_j^l are labels of fuzzy sets in U_i and V_j , respectively, $1 \leq p \leq n$, $1 \leq q \leq m$, and $l = 1, 2, \dots, M$. It is clear that R^l of (7) can be decomposed into a collection of q rules

$$R^l = [R_1^l, \dots, R_q^l] \quad (8)$$

where

$$R_j^l: \text{IF } (x_1 \text{ is } F_1^l \text{ and } \dots \text{ and } x_p \text{ is } F_p^l); \\ \text{THEN } (y_j \text{ is } G_j^l) \quad (9)$$

$j = 1, 2, \dots, q.$

A fuzzy inference engine uses the rules in the fuzzy rule base to determine a mapping from fuzzy sets in U to fuzzy sets in V based on fuzzy logic operations. In the fuzzy inference engine, the IF part of R_j^l defines a cartesian product of F_1^l, \dots, F_p^l , and the R_j^l itself is viewed as a fuzzy implication $F_1^l \times \dots \times F_p^l \rightarrow G_j^l$. Let A be an arbitrary fuzzy set in U , then each R_j^l of (9) determines a fuzzy set $A \circ R_j^l$ in V_j based on the supstar composition

$$\mu_{A \circ R_j^l}(y_j) = \sup_{\underline{x} \in U} [\mu_A(\underline{x}) \star \mu_{F_1^l \times \dots \times F_p^l \rightarrow G_j^l}(\underline{x}, y_j)] \\ = \sup_{\underline{x} \in U} [\mu_A(\underline{x}) \star \mu_{F_1^l}(x_1) \star \dots \star \mu_{F_p^l}(x_p) \star \mu_{G_j^l}(y_j)] \quad (10)$$

where $y_j \in V_j$. The final fuzzy set $A \circ R_j$ ($R_j = [R_j^1, \dots, R_j^M]$) in V_j determined by the fuzzy inference engine is obtained by combining (10) for $l = 1, 2, \dots, M$ using the triangular conorm $\dot{+}$

$$\mu_{A \circ R_j}(y_j) = \mu_{A \circ R_j^1}(y_j) \dot{+} \dots \dot{+} \mu_{A \circ R_j^M}(y_j). \quad (11)$$

In summary, the fuzzy inference engine determines a mapping from fuzzy set A in U to fuzzy set $A \circ R$ in V by¹

$$\mu_{A \circ R}(\underline{y}) = (\mu_{A \circ R_1}(y_1), \dots, \mu_{A \circ R_m}(y_m))^T. \quad (12)$$

A fuzzifier maps a crisp point $\underline{x} = (x_1, \dots, x_n)^T \in U$ into a fuzzy set $A_x = [A_{x_1}, \dots, A_{x_n}]$ in $U = U_1 \times \dots \times U_n$, where A_{x_i} is a fuzzy set in U_i . If A_{x_i} is a fuzzy singleton with support x_i , i.e., $\mu_{A_{x_i}}(x'_i) = 1$ for $x'_i = x_i$ and $\mu_{A_{x_i}}(x'_i) = 0$ for all other $x'_i \in U_i$, then we call this a singleton fuzzifier. There are other types of fuzzifiers; see [34].

A defuzzifier maps a fuzzy set in V to a crisp point in V . The defuzzifier is needed because for most practical applications; a fuzzy system is required to give a crisp output, no matter whether it is used as a controller, a decision maker, etc.. Because the output of the fuzzy inference engine is a fuzzy set $A \circ R$ in V , the defuzzifier maps $A \circ R$ into a crisp point $\underline{y} \in V$. In the literature, the most commonly used defuzzifier is the following center average defuzzifier

$$y_j = \frac{\sum_{l=1}^M \bar{y}_j^l (\mu_{A \circ R_j^l}(\bar{y}_j^l))}{\sum_{l=1}^M (\mu_{A \circ R_j^l}(\bar{y}_j^l))} \quad (13)$$

where \bar{y}_j^l is the point in V_j at which $\mu_{G_j^l}(y_j)$ achieves its maximum value, $\mu_{A \circ R_j^l}(y_j)$ is given by (10), and $j = 1, 2, \dots, m$.

We propose a modified center average defuzzifier, defined as

$$y_j = \frac{\sum_{l=1}^M \bar{y}_j^l (\mu_{A \circ R_j^l}(\bar{y}_j^l) / \sigma_j^{l2})}{\sum_{l=1}^M (\mu_{A \circ R_j^l}(\bar{y}_j^l) / \sigma_j^{l2})} \quad (14)$$

where σ_j^{l2} is a parameter characterizing the shape of $\mu_{G_j^l}(y_j)$ such that the narrower the shape of $\mu_{G_j^l}(y_j)$, the smaller is σ_j^{l2} ;

¹ We assume that each y_j for $1 \leq j \leq m$ appears at least once in the R^l of (7), therefore $\mu_{A \circ R}(\underline{y})$ is an m -dimensional vector.

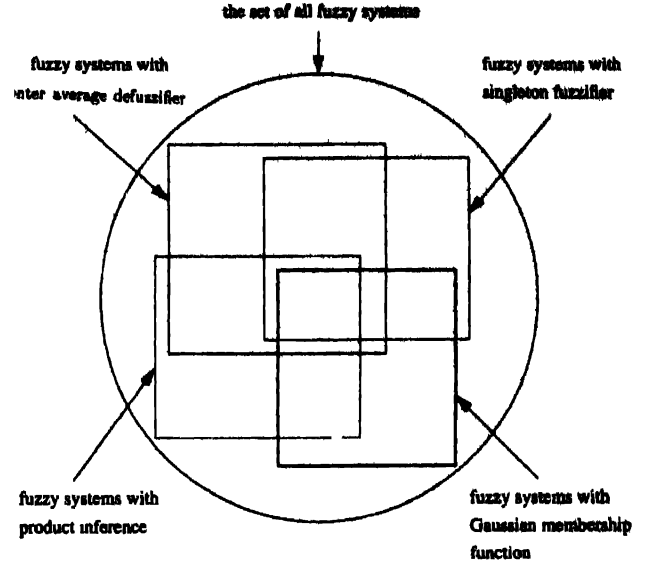


Fig. 2. Some subclasses of fuzzy systems.

for example, if $\mu_{G_j^l}(y_j) = \exp[-((y_j - \bar{y}_j^l)/\sigma_j^{l2})^2]$, then σ_j^{l2} is such a parameter. The modified center average defuzzifier is justified as follow. Common sense indicates that the sharper the shape of $\mu_{G_j^l}(y_j)$ is, the stronger is our belief that the output y_j should be nearer to $\bar{y}_j^l = \text{argsup}_{y_j \in V_j} (\mu_{G_j^l}(y_j))$ [according to the rule R_j^l of (9)]. The standard center average defuzzifier, (13), is a weighted average of the \bar{y}_j^l 's, and the weight $\mu_{A \circ R_j^l}(\bar{y}_j^l)$ determined by (10) do not take the shape of $\mu_{G_j^l}(y_j)$ into consideration. This is clearly not satisfactory based on our common sense. An obvious improvement is the modified center average defuzzifier (14).

Note that if we use the center average or modified center average defuzzifiers, we do not need to calculate the $\mu_{A \circ R_j}(y_j)$ of (11); we only need to calculate the $\mu_{A \circ R_j^l}(y_j)$ of (10) in the fuzzy inference engine.

We see that the fuzzy systems of Fig. 1 comprise a very rich class of static systems mapping from $U \subset R^n$ to $V \subset R^m$, because within each block there are many different choices, and many combinations of these choices can result in useful subclasses of fuzzy systems. If we view all the fuzzy systems as a set of functions, then each combination of these choices corresponds to a subset of the fuzzy systems, as shown for some examples in Fig. 2. We now show the specific functional forms of two subclasses of fuzzy systems which will be used later as basic building blocks of nonlinear system identifiers.

Definition 1: The set of fuzzy systems with singleton fuzzifier, center average defuzzifier, and product inference is all functions $f = (f_1, \dots, f_m)^T$ from U to V of the following form

$$f_j(\underline{x}) = \frac{\sum_{l=1}^M \bar{y}_j^l \left(\prod_{i=1}^p \mu_{F_i^l}(x_i) \right) \mu_{G_j^l}(\bar{y}_j^l)}{\sum_{l=1}^M \left(\prod_{i=1}^p \mu_{F_i^l}(x_i) \right) \mu_{G_j^l}(\bar{y}_j^l)} \quad (15)$$

where $\underline{x} = (x_1, \dots, x_n)^T \in U$, $1 \leq p \leq n$, and \bar{y}_j^l is the point at which $\mu_{G_j^l}(y_j)$ achieves its maximum value. Usually, $\mu_{G_j^l}(\bar{y}_j^l) = 1$. (15) is obtained by substituting (10) into (13),

replacing \star with the algebraic product in (1), and using the fact that if A is a fuzzy singleton with support \underline{x} (i.e., $\mu_A(\underline{x}) = 1$ and $\mu_A(\underline{x}') = 0$ for all $\underline{x}' \neq \underline{x}$), then $\mu_{A \circ R_j^i}(\underline{y}_j) = \sup_{\underline{x}' \in U} [\mu_A(\underline{x}') \star \mu_{F_1^i}(\underline{x}_1') \star \cdots \star \mu_{F_p^i}(\underline{x}_p') \star \mu_{G_j^i}(\underline{y}_j)] = \mu_{F_1^i}(\underline{x}_1) \star \cdots \star \mu_{F_p^i}(\underline{x}_p) \star \mu_{G_j^i}(\underline{y}_j)$.

Definition 2: The set of fuzzy systems with singleton fuzzifier, modified center average defuzzifier, product inference, and Gaussian membership function is all functions $f = (f_1, \dots, f_m)^T$ from U to V of the following form

$$f_j(\underline{x}) = \frac{\sum_{i=1}^M \bar{y}_j^i \left[\prod_{l=1}^p a_l^i \exp \left(-\frac{1}{2} \left(\frac{x_l - \bar{x}_l^i}{\sigma_l^i} \right)^2 \right) \right]}{\sum_{i=1}^M \left[\prod_{l=1}^p a_l^i \exp \left(-\frac{1}{2} \left(\frac{x_l - \bar{x}_l^i}{\sigma_l^i} \right)^2 \right) \right]} \bigg/ (\delta_j^i)^2 \quad (16)$$

which is obtained by taking $\mu_{F_l^i}(x_l) = a_l^i \exp(-(1/2)((x_l - \bar{x}_l^i)/\sigma_l^i)^2)$, $\mu_{G_j^i}(\underline{y}_j) = \exp(-(1/2)((\underline{y}_j - \bar{\underline{y}}_j^i)/\delta_j^i)^2)$, using the modified center average defuzzifier (14) with $\mu_{A \circ R_j^i}(\bar{\underline{y}}_j^i)$ given by (10) and $\star =$ algebraic product, and noticing that $\mu_A(\underline{x}) = 1$ and $\mu_A(\underline{x}') = 0$ for all $\underline{x}' \neq \underline{x}$, where $0 < a_l^i \leq 1$, $\sigma_l^i > 0$, $\delta_j^i > 0$, $\bar{x}_l^i \in U_l$, and $\bar{y}_j^i \in V_j$ are parameters.

The following theorem was proven in [34].

Theorem 1. Let Y be the set of fuzzy systems in Definition 2, and assume that U is compact. Then, for any real continuous function $g(\underline{x}) = (g_1(\underline{x}), \dots, g_m(\underline{x}))^T$ from U to V and arbitrary $\epsilon > 0$, there exist $f = (f_1(\underline{x}), \dots, f_m(\underline{x}))^T \in Y$ such that

$$\sup_{\underline{x} \in U} |f_j(\underline{x}) - g_j(\underline{x})| < \epsilon \quad (17)$$

for all $j = 1, 2, \dots, m$.

Next, we study the capability of using the fuzzy systems in Definition 2 to approximate dynamic systems over a finite time interval. Consider the dynamic system

$$\dot{\underline{x}} = A\underline{x} + g(\underline{x}, \underline{u}) \quad (18)$$

$$\underline{y} = h(\underline{x}, \underline{u}) \quad (19)$$

where $\underline{u} \in U$ is the input, $\underline{y} \in V$ is the output, $\underline{x} \in W \subset R^r$ is the state, A is a known Hurwitz matrix, and g, h are unknown continuous functions. We make the following assumption.

Assumption 2.1 For any \underline{u} in compact U and any finite initial condition $\underline{x}(0) = \underline{x}^0$, the solution $\underline{x}(t)$ of (18) satisfies $|\underline{x}(t) - \underline{x}^0| \leq b$ for some positive constant b and all $t \in [0, T]$, where T is an arbitrary positive constant.

Define K to be the set

$$K = \{(\underline{x}, \underline{u}) \in R^{r+n} : |\underline{x} - \underline{x}^0| \leq b + \epsilon, \underline{u} \in U\} \quad (20)$$

where $\epsilon > 0$ is an arbitrary constant, and \underline{x}^0 is any finite initial condition. Because U is compact and $|\underline{x}| \leq |\underline{x}^0| + b + \epsilon < \infty$, K is compact. Assumption 2.1 assures that for $\underline{u} \in U$ and finite \underline{x}^0 , (18) generates $(\underline{x}(t), \underline{u}(t)) \in K$ for $t \in [0, T]$. Let $\hat{g}(\underline{x}, \underline{u}|\Theta_g) = (\hat{g}_1(\underline{x}, \underline{u}|\Theta_g), \dots, \hat{g}_m(\underline{x}, \underline{u}|\Theta_g))^T$ and $\hat{h}(\underline{x}, \underline{u}|\Theta_h) = (\hat{h}_1(\underline{x}, \underline{u}|\Theta_h), \dots, \hat{h}_m(\underline{x}, \underline{u}|\Theta_h))^T$ be the fuzzy systems in Definition 2, where Θ_g, Θ_h are collections of the parameters $a_l^i, \bar{x}_l^i, \sigma_l^i$, and \bar{y}_j^i . Define Θ_g^* and Θ_h^* be such that

$$\sup_{\underline{x}, \underline{u} \in K} |\hat{g}(\underline{x}, \underline{u}|\Theta_g^*) - g(\underline{x}, \underline{u})| < \epsilon_g \quad (21)$$

$$\sup_{\underline{x}, \underline{u} \in K} |\hat{h}(\underline{x}, \underline{u}|\Theta_h^*) - h(\underline{x}, \underline{u})| < \epsilon_h \quad (22)$$

for some arbitrary $\epsilon_g > 0$ and $\epsilon_h > 0$. Theorem 1 assures the existence of the Θ_g^* and Θ_h^* . We now show that by replacing g and h in (18) and (19) with the \hat{g} and \hat{h} , we obtain a dynamic system whose output can approximate the output of (18) and (19) to arbitrary accuracy over any finite interval of time.

Theorem 2: Consider the dynamic system

$$\dot{\underline{x}} = A\underline{x} + \hat{g}(\underline{x}, \underline{u}|\Theta_g^*) \quad (23)$$

$$\underline{y} = \hat{h}(\underline{x}, \underline{u}|\Theta_h^*) \quad (24)$$

where $\underline{x}(0) = \underline{x}^0 = \underline{x}^0$, $\underline{u} \in U$, \hat{g} and \hat{h} are the fuzzy systems in Definition 2, and the parameters Θ_g^* and Θ_h^* are defined in (21) and (22). Then, for any $\epsilon > 0$, finite $T > 0$, and properly chosen ϵ_g and ϵ_h , we have that

$$\sup_{t \in [0, T]} |\underline{y}(t) - \underline{y}(t)| < \epsilon. \quad (25)$$

Proof of this theorem is given in Appendix A.

III. THE FIRST FUZZY IDENTIFIER

Consider the identification of the nonlinear system

$$\dot{\underline{x}} = f(\underline{x}) + g(\underline{x})u \quad (26)$$

where f and g are unknown functions from $U = U_1 \times \cdots \times U_n$ to $V = V_1 \times \cdots \times V_n$, $U_l, V_l \subset R$, $l = 1, 2, \dots, n$, and the input $u \in R$ and the state $\underline{x} \in R^n$ are assumed to be bounded and available for measurement. Our purpose is to develop an identification model where the f and g are replaced by fuzzy systems $\hat{f}(\underline{x}|\Theta_f)$ and $\hat{g}(\underline{x}|\Theta_g)$ and an adaptive law for updating the parameter matrices² Θ_f and Θ_g such that the identification model converges to the real system (26). First, we rewrite (26) as

$$\dot{\underline{x}} = \hat{f}(\underline{x}|\Theta_f^*) + \hat{g}(\underline{x}|\Theta_g^*)u + \underline{w} \quad (27)$$

where

$$\underline{w} \equiv [f(\underline{x}) - \hat{f}(\underline{x}|\Theta_f^*)] + [g(\underline{x}) - \hat{g}(\underline{x}|\Theta_g^*)]u, \quad (28)$$

$$\Theta_f^* \equiv \arg\min_{\Theta_f \in \Omega_f} [\sup_{\underline{x} \in U} |f(\underline{x}) - \hat{f}(\underline{x}|\Theta_f)|], \quad (29)$$

$$\Theta_g^* \equiv \arg\min_{\Theta_g \in \Omega_g} [\sup_{\underline{x} \in U} |g(\underline{x}) - \hat{g}(\underline{x}|\Theta_g)|] \quad (30)$$

and Ω_f and Ω_g are bounded feasible sets of Θ_f and Θ_g , respectively. We assume that $\Omega_f \equiv [\Theta: \text{tr}(\Theta\Theta^T) \leq M_f]$ and $\Omega_g \equiv [\Theta: \text{tr}(\Theta\Theta^T) \leq M_g]$, where M_f and M_g are given constants. Because of Theorem 1 and the boundedness of $u, |\underline{w}|$ can be arbitrarily small; therefore, without loss of generality we assume that $w_0 \equiv \sup_{t \geq 0} |\underline{w}(t)|$ is finite. We use the following serial-parallel identification model [24]

$$\dot{\underline{x}} = -\alpha \underline{x} + \alpha \underline{x} + \hat{f}(\underline{x}|\Theta_f) + \hat{g}(\underline{x}|\Theta_g)u \quad (31)$$

where α is a given positive scalar. The whole identification scheme is shown in Fig. 3. The goals of identification are the following.

²In the first fuzzy identifier we collect the parameters of the fuzzy systems into matrices, in the second fuzzy identifier we collect the parameters of the fuzzy systems into vectors

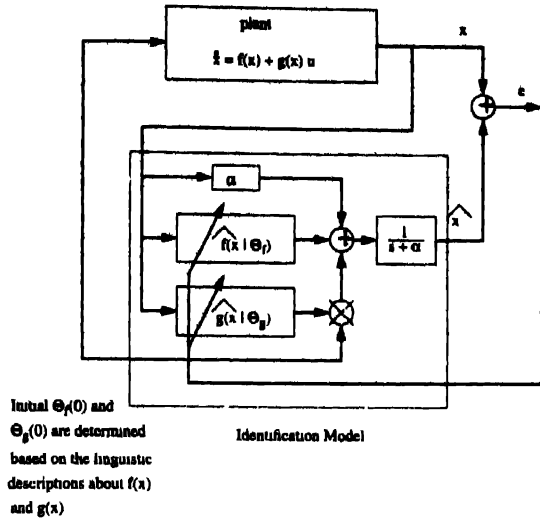


Fig. 3 Identification model using fuzzy systems.

Identification Goals: Specify the fuzzy systems $\hat{f}(\underline{x}|\Theta_f)$ and $\hat{g}(\underline{x}|\Theta_g)$, and develop an adaptive law for the parameters Θ_f and Θ_g such that:

- a) all signals involved in the identification model must be uniformly bounded, i.e., it must be guaranteed that $\hat{\underline{x}} \in L_\infty$, $\text{tr}(\Theta_f \Theta_f^T) \leq M_f$, and $\text{tr}(\Theta_g \Theta_g^T) \leq M_g$ (the input u and the system state \underline{x} are uniformly bounded by assumption), and
- b) the error $\underline{e} \equiv \underline{x} - \hat{\underline{x}}$ should be as small as possible.

To emphasize the point that our fuzzy identifiers can make use of linguistic descriptions about the unknown system (26), we make the following assumption.

Assumption 3.1: There are the following linguistic descriptions about the unknown functions $f(\underline{x})$ and $g(\underline{x})$

$$R_f^r: \text{IF } x_1 \text{ is } A_1^r \text{ and } \dots \text{ and } x_n \text{ is } A_n^r, \\ \text{THEN } f_1(\underline{x}) \text{ is } C_1^r, \dots, f_n(\underline{x}) \text{ is } C_n^r \quad (32)$$

and

$$R_g^s: \text{IF } x_1 \text{ is } B_1^s \text{ and } \dots \text{ and } x_n \text{ is } B_n^s, \\ \text{THEN } g_1(\underline{x}) \text{ is } D_1^s, \dots, g_n(\underline{x}) \text{ is } D_n^s \quad (33)$$

respectively, where A_i^r and B_i^s are fuzzy sets in U_i , C_i^r and D_i^s are fuzzy sets in V_i which achieve membership value one at some point, $r = 1, 2, \dots, N_f$, and $s = 1, 2, \dots, N_g$. We allow $N_f = N_g = 0$ which means that there are no linguistic descriptions about f and g .

Design of the First Fuzzy Identifier

Step 1: Define m_i fuzzy sets F_i^{j1} in U_i , such that for any $x_i \in U_i$, there exists at least one $\mu_{F_i^{j1}}(x_i) \neq 0$, where $i = 1, 2, \dots, n$, and $j_i = 1, 2, \dots, m_i$. We require that these F_i^{j1} 's include the A_i^r 's and B_i^s 's in (32) and (33). These fuzzy sets are fixed and will not change during the adaptation procedure of Step 4.

Step 2: Construct the fuzzy rules base of fuzzy system $\hat{f}(\underline{x}|\Theta_f)$ which consists of the following $\prod_{i=1}^n m_i$ rules

$$R^l: \text{IF } x_1 \text{ is } F_1^{j1} \text{ and } \dots \text{ and } x_n \text{ is } F_n^{jn}, \\ \text{THEN } \hat{f}_1(\underline{x}) \text{ is } G_1^l, \dots, \hat{f}_n(\underline{x}) \text{ is } G_n^l \quad (34)$$

where F_i^{j1} are defined in Step 1, G_i^l are fuzzy sets in V_i with $\mu_{G_i^l}(\theta_{if}^l) = 1$ for some parameter $\theta_{if}^l \in V_i$, $\Theta_f = (\theta_{1f}, \dots, \theta_{nf})^T$, $\theta_{if} = (\theta_{if}^1, \theta_{if}^2, \dots, \theta_{if}^{\prod_{i=1}^n m_i})^T$, $j_i = 1, 2, \dots, m_i$, $i = 1, 2, \dots, n$, and $l = 1, 2, \dots, \prod_{i=1}^n m_i$ with each l corresponds to a combination of (j_1, \dots, j_n) . Because the fuzzy rule base consists of all the possible rules concerning the fuzzy sets F_i^{j1} of Step 1 which include the A_i^r 's, it includes the linguistic descriptions in (32).

The initial parameters $\theta_{if}^l(0)$ are chosen as follow: if the IF part of G_i^l in (34) agrees with an IF part in (32), choose $\theta_{if}^l(0) = \text{argsup}_{y_i \in V_i} [\mu_{C_i^r}(y_i)]$; otherwise, choose $\theta_{if}^l(0)$ arbitrarily in V_i . Therefore, we incorporate the linguistic descriptions into the fuzzy identifier by constructing the initial fuzzy identifier based on these descriptions.

The fuzzy rule base and initial parameters of $\hat{g}(\underline{x}|\Theta_g)$ are determined in exactly the same way as $\hat{f}(\underline{x}|\Theta_f)$; we omit the details.

Step 3: Choose $\hat{f}(\underline{x}|\Theta_f) = (\hat{f}_1(\underline{x}|\theta_{1f}), \dots, \hat{f}_n(\underline{x}|\theta_{nf}))^T$ and $\hat{g}(\underline{x}|\Theta_g) = (\hat{g}_1(\underline{x}|\theta_{1g}), \dots, \hat{g}_n(\underline{x}|\theta_{ng}))^T$ to be fuzzy systems with singleton fuzzifier, center average defuzzifier, and product inference (Definition 1) based on the fuzzy rule bases constructed in Step 2

$$\hat{f}_i(\underline{x}|\theta_{if}) = \frac{\sum_{j_1=1}^{m_1} \dots \sum_{j_n=1}^{m_n} \theta_{if}^l [\mu_{F_1^{j_1}}(x_1) \dots \mu_{F_n^{j_n}}(x_n)]}{\sum_{j_1=1}^{m_1} \dots \sum_{j_n=1}^{m_n} [\mu_{F_1^{j_1}}(x_1) \dots \mu_{F_n^{j_n}}(x_n)]} \quad (35)$$

$$\hat{g}_i(\underline{x}|\theta_{ig}) = \frac{\sum_{j_1=1}^{m_1} \dots \sum_{j_n=1}^{m_n} \theta_{ig}^l [\mu_{F_1^{j_1}}(x_1) \dots \mu_{F_n^{j_n}}(x_n)]}{\sum_{j_1=1}^{m_1} \dots \sum_{j_n=1}^{m_n} [\mu_{F_1^{j_1}}(x_1) \dots \mu_{F_n^{j_n}}(x_n)]} \quad (36)$$

where $i = 1, 2, \dots, n$, and $l = 1, 2, \dots, \prod_{i=1}^n m_i$ with each l corresponds to a combination of (j_1, \dots, j_n) . Because of the way we define the fuzzy sets F_i^{j1} in Step 1, the denominators of (35) and (36) are nonzero for all $\underline{x} \in U$. Defining the fuzzy basis function [39]

$$p^l(\underline{x}) = \frac{\mu_{F_1^{j_1}}(x_1) \dots \mu_{F_n^{j_n}}(x_n)}{\sum_{j_1=1}^{m_1} \dots \sum_{j_n=1}^{m_n} [\mu_{F_1^{j_1}}(x_1) \dots \mu_{F_n^{j_n}}(x_n)]} \quad (37)$$

[l is defined as in (35) and (36)], collecting them into an $\prod_{i=1}^n m_i \times 1$ vector $\underline{p}(\underline{x})$, and collecting the θ_{if}^l and θ_{ig}^l into $\prod_{i=1}^n m_i \times 1$ vectors $\underline{\theta}_{if}$ and $\underline{\theta}_{ig}$ in the same order as $\underline{p}(\underline{x})$, we can rewrite (35) and (36) as

$$\hat{f}_i(\underline{x}|\theta_{if}) = \underline{\theta}_{if}^T \underline{p}(\underline{x}) \quad (38)$$

and

$$\hat{g}_i(\underline{x}|\theta_{ig}) = \underline{\theta}_{ig}^T \underline{p}(\underline{x}) \quad (39)$$

respectively. Recall that $\Theta_f \equiv (\theta_{1f}, \dots, \theta_{nf})^T$ and $\Theta_g \equiv (\theta_{1g}, \dots, \theta_{ng})^T$ and substitute (38) and (39) into (31); the fuzzy identifier becomes

$$\dot{\hat{\underline{x}}} = -\alpha \hat{\underline{x}} + \alpha \underline{x} + \Theta_f \underline{p}(\underline{x}) + \Theta_g \underline{p}(\underline{x}) u. \quad (40)$$

Step 4: Update the parameter matrices Θ_f and Θ_g using the following adaptive law

$$\dot{\Theta}_f = \begin{cases} \gamma_1 \underline{e} p^T(\underline{x}) & \text{if } (\text{tr}(\Theta_f \Theta_f^T) < M_f) \\ \text{or } (\text{tr}(\Theta_f \Theta_f^T) = M_f \text{ and } \underline{e}^T \Theta_f p(\underline{x}) \leq 0) \\ P[\gamma_1 \underline{e} p^T(\underline{x})] & \text{if } (\text{tr}(\Theta_f \Theta_f^T) = M_f \text{ and } \underline{e}^T \Theta_f p(\underline{x}) > 0), \end{cases} \quad (41)$$

$$\dot{\Theta}_g = \begin{cases} \gamma_2 \underline{e} p^T(\underline{x}) u & \text{if } (\text{tr}(\Theta_g \Theta_g^T) < M_g) \\ \text{or } (\text{tr}(\Theta_g \Theta_g^T) = M_g \text{ and } \underline{e}^T \Theta_g p(\underline{x}) u \leq 0) \\ P[\gamma_2 \underline{e} p^T(\underline{x}) u] & \text{if } (\text{tr}(\Theta_g \Theta_g^T) = M_g \text{ and } \underline{e}^T \Theta_g p(\underline{x}) u > 0) \end{cases} \quad (42)$$

where γ_1 and γ_2 are positive constants, $P[*]$ is the projection operator [7] defined as

$$P[\gamma_1 \underline{e} p^T(\underline{x})] \equiv \gamma_1 \underline{e} p^T(\underline{x}) - \gamma_1 \frac{\underline{e}^T \Theta_f p(\underline{x})}{\text{tr}(\Theta_f \Theta_f^T)} \Theta_f, \quad (43)$$

$$P[\gamma_2 \underline{e} p^T(\underline{x}) u] \equiv \gamma_2 \underline{e} p^T(\underline{x}) u - \gamma_2 \frac{\underline{e}^T \Theta_g p(\underline{x}) u}{\text{tr}(\Theta_g \Theta_g^T)} \Theta_g \quad (44)$$

$\underline{e} = \underline{x} - \hat{x}$, and the initial $\Theta_f(0)$ and $\Theta_g(0)$ are determined in Step 2.

Properties of the above fuzzy identifier are summarized in the following theorem.

Theorem 3: The fuzzy identifier (40) with the adaptive law (41) and (42) guarantees the following properties:

- a) $\text{tr}(\Theta_f \Theta_f^T) \leq M_f$, $\text{tr}(\Theta_g \Theta_g^T) \leq M_g$, and $\underline{\hat{x}} \in L_\infty$;
- b) there exist constants k_1 and k_2 such that

$$\int_0^t |\underline{e}(\tau)|^2 d\tau \leq k_1 + k_2 \int_0^t |\underline{w}(\tau)|^2 d\tau \quad (45)$$

for any $t \geq 0$, where $\underline{w}(\tau)$ is defined by (28); and

- c) if $\underline{w} \in L_2$, then $\lim_{t \rightarrow \infty} |\underline{e}(t)| = 0$.

Proof of this theorem is given in Appendix A.

Remark 3.1: From c) of Theorem 3 we see that to have $\lim_{t \rightarrow \infty} |\underline{e}(t)| = 0$, we require that \underline{w} is squared integrable. From the definition of \underline{w} (28)–(30), we see that \underline{w} is the sum of the minimum approximation errors of \hat{f} and \hat{g} to f and g over the entire state space U . From Theorem 1, this minimum approximation error should be small if we properly choose the fuzzy systems \hat{f} and \hat{g} .

Remark 3.2: From Step 2 we see that the linguistic descriptions (32) and (33) about the unknown functions $f(\underline{x})$ and $g(\underline{x})$ are used to construct the initial fuzzy identifier. If these descriptions provide good pictures of $f(\underline{x})$ and $g(\underline{x})$, we can expect that the adaptation procedure will converge fast, because the initial identifier constructed from good descriptions should be close to the true system. If no linguistic descriptions are available, the fuzzy identifier becomes a regular nonlinear identifier, similar to the radial basis function [3], [27] and neural network [28] identifiers.

Remark 3.3: In this fuzzy identifier, we fix the fuzzy sets in U and consider fuzzy systems whose IF parts are concerned only with these fuzzy sets. An advantage of doing so is that the fuzzy systems in the fuzzy identifier are linear in the parameter, therefore: 1) we were able to use a relatively simpler adaptive law to update the parameters, and 2) convergence of the

adaptation procedure is expected to be faster because we are not concerned with complicated nonlinear search problems. A disadvantage is that we have to consider all the possible combinations of the fuzzy sets in U , because these fuzzy sets cannot change so that we should have rules to cover every region of U , where by 'cover' we mean that for each $\underline{x} \in U$ there should exist at least one rule in the fuzzy rule bases of \hat{f} and \hat{g} whose 'strength' $(\mu_{F_1^1}(x_1) \cdots \mu_{F_n^1}(x_n))$ is not very small. Therefore, if the dimension n of a problem is large, then we have to choose the m_i 's relatively small so that the computational requirements of the fuzzy identifier do not exceed the capability of computing sources available. Clearly, the bigger the m_i 's, the more the rules, and therefore the smaller the minimum approximation error \underline{w} , which in turn means the smaller the identification error \underline{e} [see (45)]. In practical applications of this fuzzy identifier, we have to find a compromise between complexity and accuracy.

IV. THE SECOND FUZZY IDENTIFIER

As discussed in Remark 3.3, the fuzzy identifier in Section III may require a large number of rules for higher dimensional systems. An obvious way to overcome this rule explosion problem is to allow the fuzzy sets in U also to change during the adaptation procedure so that in principle any rule can cover any region of U ; therefore, we only need a small number of rules. This is the basic idea of the fuzzy identifier in this section. The price paid for this additional freedom is that the fuzzy identifier becomes nonlinear in the parameter, so that we have to use a more complicated adaptive law.

We consider the identification of the same system (26) and use the same serial-parallel model (31) (Fig. 3). The identification goals remain the same as in Section III, and we still consider Assumption 3.1. The design of the second fuzzy identifier, however, is quite different from the first one.

Design of the Second Fuzzy Identifier

Step 1. Define $2N$ fuzzy sets F_i^l and G_i^l in each U_i of U and $2N$ fuzzy sets E_i^l and H_i^l in each V_i of V with the following membership functions

$$\mu_{F_i^l}(x_i) = \exp \left[-\frac{1}{2} \frac{(x_i - \bar{x}_{if}^l)^2}{(\sigma_{if}^l)^2 + \epsilon} \right] \quad (46)$$

$$\mu_{G_i^l}(x_i) = \exp \left[-\frac{1}{2} \frac{(x_i - \bar{x}_{ig}^l)^2}{(\sigma_{ig}^l)^2 + \epsilon} \right] \quad (47)$$

$$\mu_{E_i^l}(y_i) = \exp \left[-\frac{1}{2} \frac{(y_i - \bar{y}_{ie}^l)^2}{(\sigma_{ie}^l)^2 + \epsilon} \right] \quad (48)$$

$$\mu_{H_i^l}(y_i) = \exp \left[-\frac{1}{2} \frac{(y_i - \bar{y}_{ih}^l)^2}{(\sigma_{ih}^l)^2 + \epsilon} \right] \quad (49)$$

where $\epsilon > 0$ is a small constant, $l = 1, 2, \dots, N$ (in general, $N \ll \prod_{i=1}^n m_i$), $x_i \in U_i$, $y_i \in V_i$, $i = 1, 2, \dots, n$, and \bar{x}_{if}^l , σ_{if}^l , \bar{x}_{ig}^l , σ_{ig}^l , \bar{y}_{ie}^l , σ_{ie}^l , \bar{y}_{ih}^l , and σ_{ih}^l are free parameters which will be updated in the adaptation procedure of Step 4. The purpose of adding the small $\epsilon > 0$ to the fuzzy membership functions (46)–(49) is that even if the σ 's = 0, the fuzzy membership functions are still well defined. This modification

will make the adaptive law simpler because we do not require the σ 's $\neq 0$. For this fuzzy identifier, we assume that the membership functions of A_i^r , B_i^s , C_i^r , and D_i^s in (32) and (33) are in the form of (46)–(49), respectively, and that $N \geq N_f$ and $N \geq N_g$.

Step 2: Construct the fuzzy rule bases of \hat{f} and \hat{g} as

$$R_f^l: \text{IF } x_1 \text{ is } F_1^l \text{ and } \dots \text{ and } x_n \text{ is } F_n^l, \\ \text{THEN } \hat{f}_1(\underline{x}) \text{ is } E_1^l, \dots, \hat{f}_n(\underline{x}) \text{ is } E_n^l \quad (50)$$

and

$$R_g^l: \text{IF } x_1 \text{ is } G_1^l \text{ and } \dots \text{ and } x_n \text{ is } G_n^l, \\ \text{THEN } \hat{g}_1(\underline{x}) \text{ is } H_1^l, \dots, \hat{g}_n(\underline{x}) \text{ is } H_n^l \quad (51)$$

respectively, where F_i^l , E_i^l , G_i^l , and H_i^l are defined in Step 1, and $l = 1, 2, \dots, N$. Choose $\hat{f}(\underline{x}|\underline{\theta}_f) = (\hat{f}_1(\underline{x}|\underline{\theta}_f), \dots, \hat{f}_n(\underline{x}|\underline{\theta}_f))^T$ and $\hat{g}(\underline{x}|\underline{\theta}_g) = (\hat{g}_1(\underline{x}|\underline{\theta}_g), \dots, \hat{g}_n(\underline{x}|\underline{\theta}_g))^T$ to be fuzzy systems with singleton fuzzifier, modified center average defuzzifier, product inference, and Gaussian membership function (Definition 2) based on the fuzzy rule bases of (50) and (51), respectively

$$\hat{f}_i(\underline{x}|\underline{\theta}_f) = \frac{\sum_{l=1}^N \bar{y}_{ie}^l \left[\prod_{j=1}^n \exp \left(-\frac{1}{2} \frac{(x_j - \bar{x}_{j,f}^l)^2}{(\sigma_{j,f}^l)^2 + \epsilon} \right) \right]}{\sum_{l=1}^N \left[\prod_{j=1}^n \exp \left(-\frac{1}{2} \frac{(x_j - \bar{x}_{j,f}^l)^2}{(\sigma_{j,f}^l)^2 + \epsilon} \right) \right]} \frac{[(\sigma_{ie}^l)^2 + \epsilon]}{[(\sigma_{ie}^l)^2 + \epsilon]}, \quad (52)$$

$$\hat{g}_i(\underline{x}|\underline{\theta}_g) = \frac{\sum_{l=1}^N \bar{y}_{ih}^l \left[\prod_{j=1}^n \exp \left(-\frac{1}{2} \frac{(x_j - \bar{x}_{j,g}^l)^2}{(\sigma_{j,g}^l)^2 + \epsilon} \right) \right]}{\sum_{l=1}^N \left[\prod_{j=1}^n \exp \left(-\frac{1}{2} \frac{(x_j - \bar{x}_{j,g}^l)^2}{(\sigma_{j,g}^l)^2 + \epsilon} \right) \right]} \frac{[(\sigma_{ih}^l)^2 + \epsilon]}{[(\sigma_{ih}^l)^2 + \epsilon]}, \quad (53)$$

where $\underline{\theta}_f$ and $\underline{\theta}_g$ are $4nN \times 1$ vectors which are collections of the free parameters. Specifically, $\underline{\theta}_f \equiv (\bar{y}_{1e}^1, \dots, \bar{y}_{1e}^N, \dots, \bar{y}_{ne}^1, \dots, \bar{y}_{ne}^N; \sigma_{1e}^1, \dots, \sigma_{1e}^N, \dots, \sigma_{ne}^1, \dots, \sigma_{ne}^N; \bar{x}_{1f}^1, \dots, \bar{x}_{1f}^N, \dots, \bar{x}_{nf}^1, \dots, \bar{x}_{nf}^N; \sigma_{1f}^1, \dots, \sigma_{1f}^N, \dots, \sigma_{nf}^1, \dots, \sigma_{nf}^N)$, and $\underline{\theta}_g$ is defined in exactly the same way as $\underline{\theta}_f$ with the subscript "e" replaced by "h" and "f" replaced by "g". Different from the first fuzzy identifier where we collected the free parameters into matrices Θ_f and Θ_g , we collect the free parameters into vectors $\underline{\theta}_f$ and $\underline{\theta}_g$ in this second fuzzy identifier. We still use the serial-parallel identification model (31), with $\hat{f}(\underline{x}|\Theta_f)$ and $\hat{g}(\underline{x}|\Theta_g)$ replaced by the $\hat{f}(\underline{x}|\underline{\theta}_f)$ and $\hat{g}(\underline{x}|\underline{\theta}_g)$ of (52) and (53), respectively.

Step 3: Compute the gradient matrices $(\partial \hat{f}(\underline{x}|\underline{\theta}_f)/\partial \underline{\theta}_f) = ((\partial \hat{f}_1(\underline{x}|\underline{\theta}_f)/\partial \underline{\theta}_f), \dots, (\partial \hat{f}_n(\underline{x}|\underline{\theta}_f)/\partial \underline{\theta}_f))$ and $(\partial \hat{g}(\underline{x}|\underline{\theta}_g)/\partial \underline{\theta}_g) = ((\partial \hat{g}_1(\underline{x}|\underline{\theta}_g)/\partial \underline{\theta}_g), \dots, (\partial \hat{g}_n(\underline{x}|\underline{\theta}_g)/\partial \underline{\theta}_g))$ using the following back-propagation algorithm which is derived in Appendix B

$$\frac{\partial \hat{f}_i}{\partial \sigma_{ie}^l} = \frac{\bar{y}_{ie}^l - \hat{f}_i}{b_i} \frac{-2z^l \sigma_{ie}^l}{((\sigma_{ie}^l)^2 + \epsilon)}, \quad (54)$$

$$\frac{\partial \hat{f}_i}{\partial \bar{x}_{j,f}^l} = \frac{\bar{y}_{ie}^l - \hat{f}_i}{b_i} \frac{1}{(\sigma_{ie}^l)^2 + \epsilon} z^l \frac{x_j - \bar{x}_{j,f}^l}{(\sigma_{j,f}^l)^2 + \epsilon}, \quad (55)$$

$$\frac{\partial \hat{f}_i}{\partial \sigma_{j,f}^l} = \frac{\bar{y}_{ie}^l - \hat{f}_i}{b_i} \frac{1}{(\sigma_{ie}^l)^2 + \epsilon} z^l \frac{2(x_j - \bar{x}_{j,f}^l)^2 \sigma_{j,f}^l}{((\sigma_{j,f}^l)^2 + \epsilon)^2}, \quad (56)$$

$$\frac{\partial \hat{f}_i}{\partial \sigma_{j,f}^l} = \frac{\bar{y}_{ie}^l - \hat{f}_i}{b_i} \frac{1}{(\sigma_{ie}^l)^2 + \epsilon} z^l \frac{2(x_j - \bar{x}_{j,f}^l)^2 \sigma_{j,f}^l}{((\sigma_{j,f}^l)^2 + \epsilon)^2}, \quad (57)$$

where

$$z^l \equiv \prod_{j=1}^n \exp \left(-\frac{1}{2} \frac{(x_j - \bar{x}_{j,f}^l)^2}{(\sigma_{j,f}^l)^2 + \epsilon} \right), \quad (58)$$

$$b_i \equiv \sum_{l=1}^N z^l / ((\sigma_{ie}^l)^2 + \epsilon) \quad (59)$$

$l = 1, 2, \dots, N$, and $i, j = 1, 2, \dots, n$; and $(\partial \hat{g}_i/\partial \bar{y}_{ih}^l)$, $(\partial \hat{g}_i/\partial \sigma_{ih}^l)$, $(\partial \hat{g}_i/\partial \bar{x}_{j,g}^l)$, and $(\partial \hat{g}_i/\partial \sigma_{j,g}^l)$ are computed in exactly the same way as (54)–(57), respectively, with the subscripts "e" replaced by "h" and "f" replaced by "g."

Step 4: Update the parameter vectors $\underline{\theta}_f$ and $\underline{\theta}_g$ using the following adaptive law

$$\underline{\dot{\theta}}_f = P \left[\gamma_1 \frac{\partial \hat{f}(\underline{x}|\underline{\theta}_f)}{\partial \underline{\theta}_f} \underline{\epsilon} \text{ if } (|\underline{\theta}_f|^2 < M_f) \right. \\ \left. \text{or } (|\underline{\theta}_f|^2 = M_f \text{ and } \underline{\epsilon}^T \left(\frac{\partial \hat{f}}{\partial \underline{\theta}_f} \right)^T \underline{\theta}_f \leq 0) \right] \\ \text{if } (|\underline{\theta}_f|^2 = M_f \text{ and } \underline{\epsilon}^T \left(\frac{\partial \hat{f}}{\partial \underline{\theta}_f} \right)^T \underline{\theta}_f > 0), \quad (60)$$

$$\underline{\dot{\theta}}_g = P \left[\gamma_2 \frac{\partial \hat{g}(\underline{x}|\underline{\theta}_g)}{\partial \underline{\theta}_g} \underline{\epsilon} u \text{ if } (|\underline{\theta}_g|^2 < M_g) \right. \\ \left. \text{or } (|\underline{\theta}_g|^2 = M_g \text{ and } \underline{\epsilon}^T \left(\frac{\partial \hat{g}}{\partial \underline{\theta}_g} \right)^T \underline{\theta}_g u \leq 0) \right] \\ \text{if } (|\underline{\theta}_g|^2 = M_g \text{ and } \underline{\epsilon}^T \left(\frac{\partial \hat{g}}{\partial \underline{\theta}_g} \right)^T \underline{\theta}_g u > 0) \quad (61)$$

where γ_1 and γ_2 are positive constants, $P[\ast]$ is the projection operator [7]

$$\gamma_1 \frac{\partial \hat{f}}{\partial \underline{\theta}_f} \underline{\epsilon} \equiv \gamma_1 \frac{\partial \hat{f}}{\partial \underline{\theta}_f} \underline{\epsilon} - \gamma_1 \frac{\underline{\epsilon}^T \left(\frac{\partial \hat{f}}{\partial \underline{\theta}_f} \right)^T \underline{\theta}_f}{|\underline{\theta}_f|^2} \underline{\theta}_f, \quad (62)$$

$$P[\gamma_2 \frac{\partial \hat{g}}{\partial \underline{\theta}_g} \underline{\epsilon} u] \equiv \gamma_2 \frac{\partial \hat{g}}{\partial \underline{\theta}_g} \underline{\epsilon} u - \gamma_2 \frac{\underline{\epsilon}^T \left(\frac{\partial \hat{g}}{\partial \underline{\theta}_g} \right)^T \underline{\theta}_g u}{|\underline{\theta}_g|^2} \underline{\theta}_g \quad (63)$$

$\underline{\epsilon} = \underline{x} - \hat{\underline{x}}$, and the initial parameters $\underline{\theta}_f(0)$ and $\underline{\theta}_g(0)$ are determined in the following way: for rules R_f^l and R_g^l in (50) and (51) that agree with the linguistic rules (32) and (33), set their initial parameters equal to the corresponding parameters in the linguistic rules; otherwise, set the initial parameters arbitrarily. Therefore, same as the first fuzzy identifier, we incorporate linguistic descriptions about f and g into the fuzzy identifier by constructing the initial fuzzy identifier based on these linguistic descriptions.

The following theorem summarizes the properties of the second fuzzy identifier.

Theorem 4: The fuzzy identifier (31) with \hat{f} and \hat{g} given by (52) and (53) and the adaptive law (60) and (61) guarantees the following properties:

- a) $|\underline{\theta}_f|^2 \leq M_f$, $|\underline{\theta}_g|^2 \leq M_g$, and $\dot{x} \in L_\infty$;
- b) there exist constants k_1 and k_2 such that

$$\int_0^t |\underline{e}(\tau)|^2 d\tau \leq k_1 + k_2 \int_0^t |\underline{v}(\tau)|^2 d\tau \quad (64)$$

where

$$\begin{aligned} \underline{v} \equiv & \left[\frac{\partial \hat{f}(x|\underline{\theta}_{f0})}{\partial \underline{\theta}_f} - \frac{\partial \hat{f}(x|\underline{\theta}_f^*)}{\partial \underline{\theta}_f} \right]^T (\underline{\theta}_f^* - \underline{\theta}_f) \\ & + \left[\frac{\partial \hat{g}(x|\underline{\theta}_{g0})}{\partial \underline{\theta}_g} - \frac{\partial \hat{g}(x|\underline{\theta}_g^*)}{\partial \underline{\theta}_g} \right]^T (\underline{\theta}_g^* - \underline{\theta}_g) u + \underline{w} \end{aligned} \quad (65)$$

$\underline{\theta}_{f0} = \lambda_1 \underline{\theta}_f + (1 - \lambda_1) \underline{\theta}_f^*$, $\underline{\theta}_{g0} = \lambda_2 \underline{\theta}_g + (1 - \lambda_2) \underline{\theta}_g^*$ for some $\lambda_1, \lambda_2 \in [0, 1]$, and $\underline{w}, \underline{\theta}_f^*, \underline{\theta}_g^*$ are defined in (28)–(30) with Θ_f^* and Θ_g^* replaced by $\underline{\theta}_f^*$ and $\underline{\theta}_g^*$;

- c) if $\underline{v} \in L_2$, then $\lim_{t \rightarrow \infty} |\underline{e}(t)| = 0$.

Proof of this theorem is given in Appendix A.

Remark 4.1: Because we have much more freedom in choosing the parameters of this fuzzy identifier than the first fuzzy identifier, the \underline{w} for this fuzzy identifier should be smaller than the first one. Therefore, the key factor that influence the value of \underline{v} is the closeness of the $\underline{\theta}_f$ and $\underline{\theta}_g$ to their optimal values $\underline{\theta}_f^*$ and $\underline{\theta}_g^*$. From (65) we see that if $|\underline{\theta}_f^* - \underline{\theta}_f|$ and $|\underline{\theta}_g^* - \underline{\theta}_g|$ are small, then \underline{v} will be small. Because the initial parameters $\underline{\theta}_f(0)$ and $\underline{\theta}_g(0)$ are determined based on the linguistic descriptions about f and g , good descriptions should make $\underline{\theta}_f(0)$ and $\underline{\theta}_g(0)$ close to $\underline{\theta}_f^*$ and $\underline{\theta}_g^*$ and therefore make \underline{v} small.

Remark 4.2: By using the adaptive law (60) and (61), it is possible that some σ parameters are zero at certain points. To deal with this problem, we added a small $\epsilon > 0$ to the fuzzy systems (52) and (53) so that the fuzzy identifier is well defined for zero σ parameters.

V. SIMULATIONS

In this section, we simulate the two fuzzy identifiers for the following chaotic glycolytic oscillator [10]

$$\dot{x}_1(t) = -x_1(t)x_2^2(t) + 0.999 + 0.42 \cos(1.75t) \quad (66)$$

$$\dot{x}_2(t) = x_1(t)x_2^2(t) - x_2(t). \quad (67)$$

Fig. 4 shows the trajectory of this chaotic system in the phase plane from $t = 0$ to $t = 50$ with initial condition $x_1(0) = x_2(0) = 1.5$. For this system, we have $f_1(x) = -x_1x_2^2 + 0.999$, $f_2(x) = x_1x_2^2 - x_2$, $g_1(x) = 0.42$, and $g_2(x) = 0$. For simplicity, we assume that g_1 and g_2 are known, i.e., we only consider the identification of f_1 and f_2 . For each fuzzy identifier, we simulate two cases: 1) without linguistic descriptions about f_1 and f_2 , and 2) with the following linguistic descriptions

R_f^1 : IF x_1 is “near 1” and x_2 is “near 1”,

THEN f_1 is “near 0”. f_2 is “near 0”, (68)

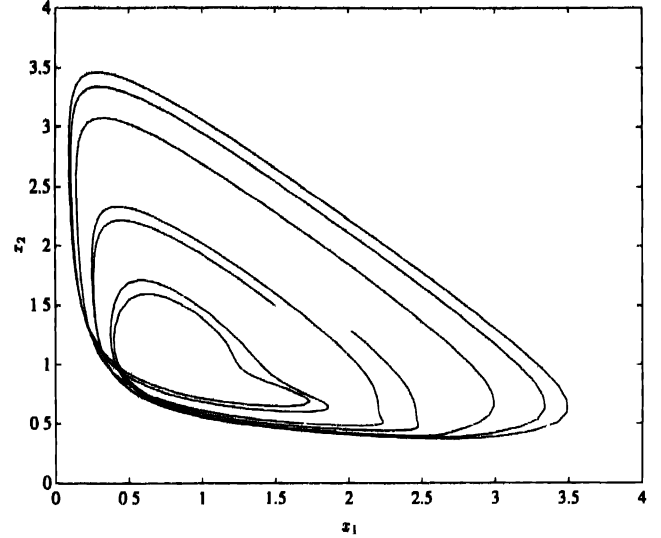


Fig. 4 Trajectory of the chaotic glycolytic oscillator in the phase plane from $t = 0$ to $t = 50$ with initial condition $x_1(0) = x_2(0) = 1.5$

R_f^2 : IF x_1 is “near 1” and x_2 is “near 2”,

THEN f_1 is “near -3”, f_2 is “near 2”, (69)

R_f^3 : IF x_1 is “near 2” and x_2 is “near 1”,

THEN f_1 is “near -1”, f_2 is “near 1” (70)

where “near c ”, $c = 1, 0, 2, -3$, or -1 , is a fuzzy set μ_c with membership function $\mu_c(z) = \exp[-(1/2)((z - c)/0.25)^2]$. $R_f^1 - R_f^3$ are obtained by evaluating f_1 and f_2 at three points $\underline{x} = (1, 1)^T$, $(1, 2)^T$, and $(2, 1)^T$, and then fuzzifying the $\underline{x} - f(\underline{x})$ pairs. For practical systems, this kind of information is provided by human experts who are familiar with the behavior of the systems. In all the simulations in the sequel, we chose $\alpha = 1$, $\gamma_1 = 4$, $M_f = 10^6$, and $\epsilon = 10^{-6}$, and solved the differential equations using the MATLAB command “ode23” which uses the second/third order Runge-Kutta method.

To simulate the first fuzzy identifier, we first need to define some fuzzy sets to cover the state space. From Fig. 4 we see that most values of x_1 and x_2 are in the interval $[0.5, 3.5]$, therefore we define the following seven fuzzy sets F^j ($j = 1, 2, \dots, 7$) for both x_1 and x_2 (i.e., $m_1 = m_2 = 7$): $\mu_{F^j}(x_i) = \exp[-(1/2)((x_i - \bar{x}^j)/0.25)^2]$, where $\bar{x}^j = 0.5 \times j$, $j = 1, 2, \dots, 7$, and $i = 1, 2$. Clearly, these fuzzy sets include the fuzzy sets in the IF parts of $R_f^1 - R_f^3$. For this choice of F^j , we have 49 fuzzy rules to construct each \hat{f}_i ($i = 1, 2$), i.e., each \hat{f}_i has 49 free parameters. We simulated two cases: 1) without incorporating the $R_f^1 - R_f^3$, i.e., all elements of $\Theta_f(0)$ were chosen randomly in the interval $[-2, 2]$; and 2) incorporating the $R_f^1 - R_f^3$, i.e., the elements of $\Theta_f(0)$ that correspond to the conditions in the IF parts of $R_f^1 - R_f^3$ were chosen according the corresponding THEN parts of $R_f^1 - R_f^3$, and other elements were still chosen randomly in the interval $[-2, 2]$. Figs. 5 and 6 show the states $x_1(t)$ and $x_2(t)$ with the corresponding $\hat{x}_1(t)$ and $\hat{x}_2(t)$, respectively, for the first case, and Figs. 7 and 8 show the same results for the second case. Comparing Figs. 5 and 7 and Figs. 6 and 8, we see that

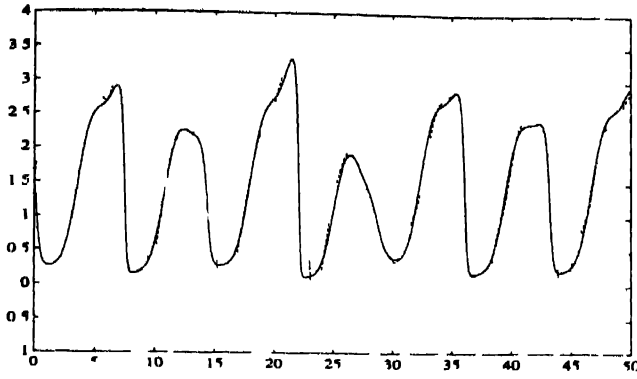


Fig 5 $r_1(t)$ (solid line) and $r_1(t)$ (dashed line) using the first fuzzy identifier without incorporating the linguistic descriptions (68)–(70)

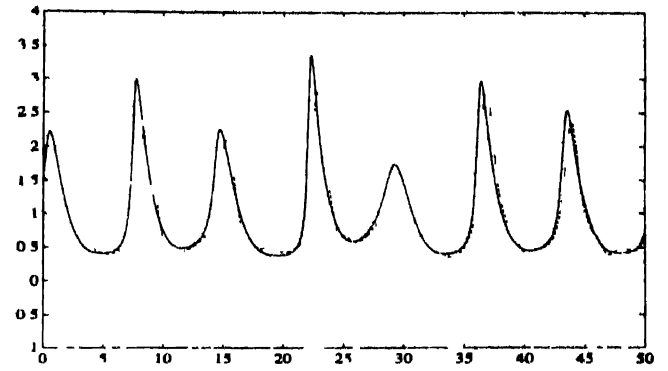


Fig 8 $r_2(t)$ (solid line) and $r_2(t)$ (dashed line) using the first fuzzy identifier after incorporating the linguistic descriptions (68)–(70)

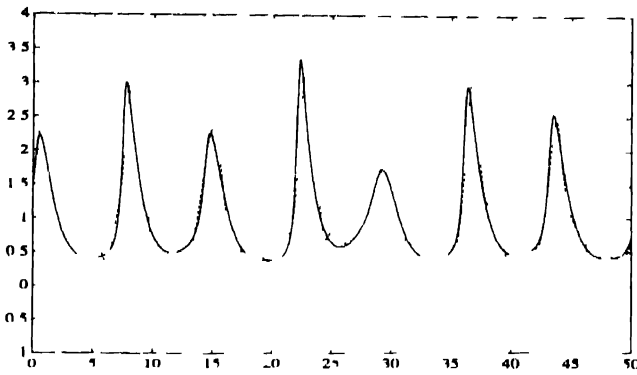


Fig 6 $r_2(t)$ (solid line) and $r_2(t)$ (dashed line) using the first fuzzy identifier without incorporating the linguistic descriptions (68)–(70)

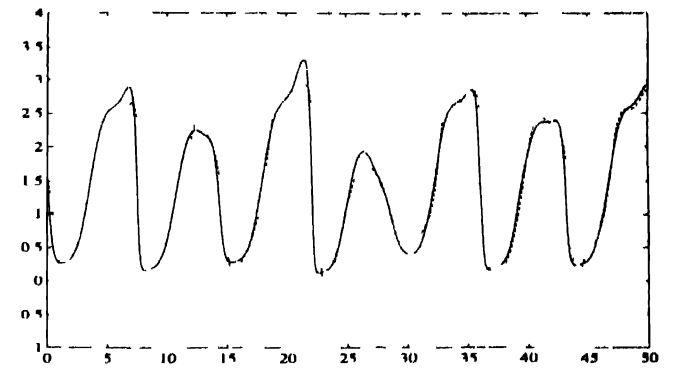


Fig 9 $r_1(t)$ (solid line) and $r_1(t)$ (dashed line) using the second fuzzy identifier without incorporating the linguistic descriptions (68)–(70)

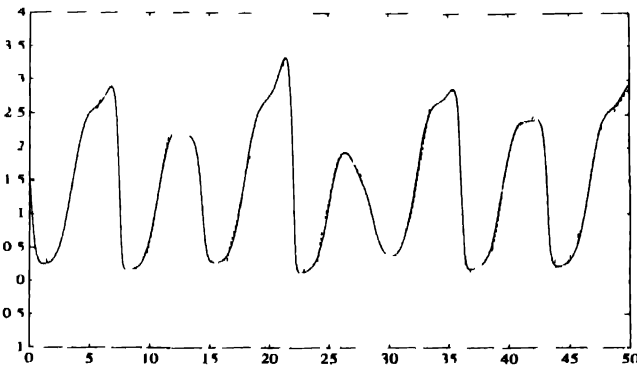


Fig 7 $r_1(t)$ (solid line) and $r_1(t)$ (dashed line) using the first fuzzy identifier after incorporating the linguistic descriptions (68)–(70)

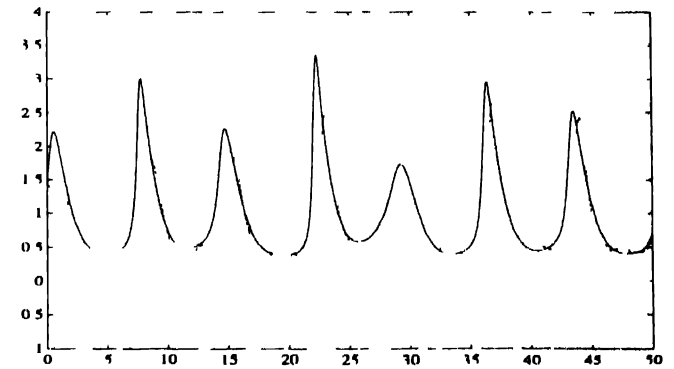


Fig 10 $r_2(t)$ (solid line) and $r_2(t)$ (dashed line) using the second fuzzy identifier without incorporating the linguistic descriptions (68)–(70)

the adaptation speed and accuracy were greatly improved by incorporating the linguistic rules $R_f^1 - R_f^3$.

We simulated the second fuzzy identifier for the same two cases: 1) without incorporating the $R_f^1 - R_f^3$, i.e., all elements of $\theta_f(0)$ were chosen randomly in the interval $[-2, 2]$, and 2) incorporating the $R_f^1 - R_f^3$, i.e., some elements of $\theta_f(0)$ were chosen according to the fuzzy sets in $R_f^1 - R_f^3$, and the remaining elements were chosen randomly in the interval $[-2, 2]$. We chose $N = 10$, so that on average each \hat{f}_i ($i = 1, 2$) contains 40 free parameters [see (52)]. Figures 9, and 10 show the $r_1(t)$, $\hat{r}_2(t)$, and $x_2(t)$, $\hat{x}_2(t)$ curves for the first case, respectively, and Figs 11 and 12 show the same

results for the second case. Comparing Figs 9 and 11 and Figs 10 and 12, we see that the adaptation speed and accuracy were greatly improved by incorporating the linguistic rules $R_f^1 - R_f^3$.

Comparing Figs. 5–8 with the corresponding Figs. 9–12, we see that although the second fuzzy identifier used less free parameters than the first fuzzy identifier, its performance is better. This may be caused by two reasons. 1) although each fuzzy system in the first fuzzy identifier has 49 free parameters, some of them were never used because from Fig. 4 we see that the states of the system never enter some regions in the state space so that the parameters responsible for these regions

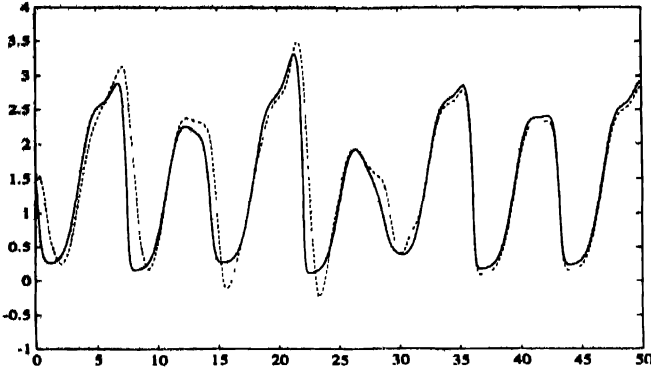


Fig. 11. $x_1(t)$ (solid line) and $\hat{x}_1(t)$ (dashed line) using the second fuzzy identifier after incorporating the linguistic descriptions (68)–(70)

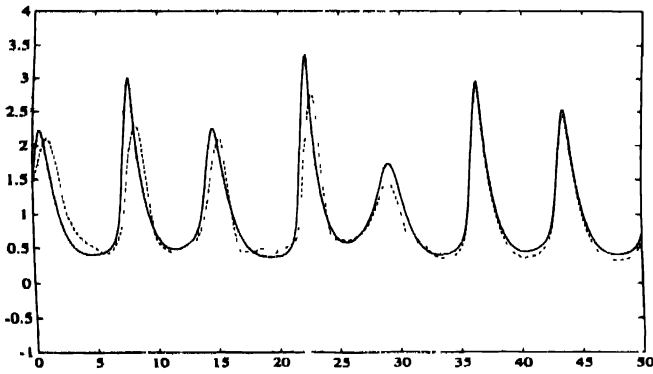


Fig. 12. $x_2(t)$ (solid line) and $\hat{x}_2(t)$ (dashed line) using the second fuzzy identifier after incorporating the linguistic descriptions (68)–(70).

have no influence on the performance of the identifier, i.e., the number of actually useful free parameters of the second fuzzy identifier is more than that of the first fuzzy identifier, and 2) the first fuzzy identifier is linear in the parameter so that in principle it should converge to a single solution, whereas the second fuzzy identifier is nonlinear in the parameter so that it has a larger functional space to search. In principle, the second fuzzy identifier has the danger to be trapped at a local minimum; but this seems not happen in our simulations, especially when we incorporated the linguistic rules.

VI. CONCLUSIONS

In this paper two identifiers of nonlinear dynamic systems were developed based on the fuzzy system models. We proved that all signals in the fuzzy identifiers are uniformly bounded and provided conditions under which the identification errors converge to zero. We also proved that the fuzzy identifiers are capable of following the output of a very general nonlinear dynamic system to arbitrary accuracy in any finite time interval. The most important advantage of the fuzzy identifiers is that linguistic descriptions about the systems (in terms of fuzzy IF-THEN rules) can be directly incorporated into the fuzzy identifiers. We simulated the two fuzzy identifiers for the chaotic glycolytic oscillator, and the results show that: 1) they could identify the chaotic system at a reasonable speed and accuracy without using any linguistic descriptions, and 2)

by incorporating some linguistic descriptions, the speed and accuracy of the fuzzy identifiers were greatly improved.

APPENDIX A

Proof of Theorem 2: From (16) we see that \hat{g} and \hat{h} are continuous functions and therefore satisfy the Lipschitz condition in the compact set K of (20), i.e., there exist constants b_g and b_h such that for all $(\underline{x}^{(1)}, \underline{u}), (\underline{x}^{(2)}, \underline{u}) \in K$

$$|\hat{g}(\underline{x}^{(1)}, \underline{u}|\Theta_g) - \hat{g}(\underline{x}^{(2)}, \underline{u}|\Theta_g)| \leq b_g |\underline{x}^{(1)} - \underline{x}^{(2)}|, \quad (\text{A.1})$$

$$|\hat{h}(\underline{x}^{(1)}, \underline{u}|\Theta_h) - \hat{h}(\underline{x}^{(2)}, \underline{u}|\Theta_h)| \leq b_h |\underline{x}^{(1)} - \underline{x}^{(2)}|. \quad (\text{A.2})$$

Define $\underline{e}_x \equiv \hat{\underline{x}} - \underline{x}$, then from (18) and (23) we have

$$\dot{\underline{e}}_x = A\underline{e}_x + \hat{g}(\hat{\underline{x}}, \underline{u}|\Theta_g^*) - g(\underline{x}, \underline{u}) \quad (\text{A.3})$$

whose solution can be expressed as

$$\underline{e}_x(t) = \int_0^t e^{A(t-\tau)} [\hat{g}(\hat{\underline{x}}(\tau), \underline{u}(\tau)|\Theta_g^*) - g(\underline{x}(\tau), \underline{u}(\tau))] d\tau. \quad (\text{A.4})$$

Since A is a Hurwitz matrix, there exist positive constants c and α such that $\|e^{At}\| \leq ce^{-\alpha t}$ for all $t \geq 0$. Let $\epsilon_g = \frac{c\alpha}{2cb_h} e^{-cb_g/\alpha}$, then from (A.4), (A.1), and (21) we have that

$$\begin{aligned} |\underline{e}_x(t)| &\leq \int_0^t \|e^{A(t-\tau)}\| |\hat{g}(\hat{\underline{x}}(\tau), \underline{u}(\tau)|\Theta_g^*) - \hat{g}(\underline{x}(\tau), \underline{u}(\tau)|\Theta_g^*)| d\tau \\ &\quad + \int_0^t \|e^{A(t-\tau)}\| |\hat{g}(\underline{x}(\tau), \underline{u}(\tau)|\Theta_g^*) - g(\underline{x}(\tau), \underline{u}(\tau))| d\tau \\ &\leq \int_0^t ce^{-\alpha(t-\tau)} b_g |\underline{e}_x(\tau)| d\tau \\ &\quad + \int_0^t ce^{-\alpha(t-\tau)} \frac{c\alpha}{2cb_h} e^{-cb_g/\alpha} d\tau \\ &\leq cb_g \int_0^t e^{-\alpha(t-\tau)} |\underline{e}_x(\tau)| d\tau + \frac{\epsilon}{2b_h} e^{-cb_g/\alpha}. \end{aligned} \quad (\text{A.5})$$

Using the Bellman–Gronwall Lemma [9], we obtain

$$\begin{aligned} |\underline{e}_x(t)| &\leq \frac{\epsilon}{2b_h} e^{-cb_g/\alpha} e^{cb_g} \int_0^t e^{-\alpha(t-\tau)} d\tau \\ &\leq \frac{\epsilon}{2b_h} e^{-cb_g/\alpha} [1 + \int_0^t e^{-\alpha(t-\tau)} d\alpha(t-\tau)] \leq \frac{\epsilon}{2b_h}. \end{aligned} \quad (\text{A.6})$$

Therefore, $|\hat{\underline{x}} - \underline{x}^0| \leq |\hat{\underline{x}} - \underline{x}| + |\underline{x} - \underline{x}^0| \leq (\epsilon/2b_h) + b$ from Assumption 2.1. Without loss of generality, we assume that $b_h \geq 1$; therefore, $(\hat{\underline{x}}, \underline{u}) \in K$. Hence, letting $\epsilon_h = \epsilon/2$, using (22), (A.2) and (A.6), and considering the fact that $(\underline{x}, \underline{u}) \in K$ and $(\hat{\underline{x}}, \underline{u}) \in K$, we obtain

$$\begin{aligned} |\hat{\underline{y}}(t) - \underline{y}(t)| &\leq |\hat{h}(\hat{\underline{x}}, \underline{u}|\Theta_h^*) - \hat{h}(\underline{x}, \underline{u}|\Theta_h^*)| \\ &\quad + |\hat{h}(\underline{x}, \underline{u}|\Theta_h^*) - h(\underline{x}, \underline{u})| \\ &\leq b_h |\underline{e}_x(t)| + \epsilon/2 \leq \epsilon \end{aligned} \quad (\text{A.7})$$

for all $t \in [0, T]$.

Q.E.D.

Proof of Theorem 3: a) From (41) and (43) we see that if $\text{tr}(\Theta_f \Theta_f^T) = M_f$, then: if $\underline{e}^T \Theta_f \underline{p}(\underline{x}) \leq 0$, we have

$$\begin{aligned} \frac{d}{dt}[\text{tr}(\Theta_f \Theta_f^T)] &= \text{tr}(\dot{\Theta}_f \Theta_f^T + \Theta_f \dot{\Theta}_f^T) \\ &= \text{tr}(\gamma_1 \underline{e} \underline{p}^T(\underline{x}) \Theta_f^T + \gamma_1 \Theta_f \underline{p}(\underline{x}) \underline{e}^T) \\ &= 2\gamma_1 \underline{e}^T \Theta_f \underline{p}(\underline{x}) \leq 0 \end{aligned} \quad (\text{A.8})$$

where we use the properties of trace in the last equality; and if $\underline{e}^T \Theta_f \underline{p}(\underline{x}) > 0$, we have

$$\begin{aligned} \frac{d}{dt}[\text{tr}(\Theta_f \Theta_f^T)] &= 2 \text{tr} \left(\gamma_1 \underline{e} \underline{p}^T(\underline{x}) \Theta_f^T - \gamma_1 \frac{\underline{e}^T \Theta_f \underline{p}(\underline{x})}{\text{tr}(\Theta_f \Theta_f^T)} \Theta_f \Theta_f^T \right) \\ &= 2\gamma_1 \underline{e}^T \Theta_f \underline{p}(\underline{x}) - 2\gamma_1 \underline{e}^T \Theta_f \underline{p}(\underline{x}) = 0. \end{aligned} \quad (\text{A.9})$$

Hence, we always have $\text{tr}(\Theta_f \Theta_f^T) \leq M_f$. Similarly, we can prove that $\text{tr}(\Theta_g \Theta_g^T) \leq M_g$. To prove $\underline{\hat{x}} \in L_\infty$, define the Lyapunov function candidate

$$V = \frac{1}{2} |\underline{e}|^2 + \frac{1}{2\gamma_1} \text{tr}(\Phi_f \Phi_f^T) + \frac{1}{2\gamma_2} \text{tr}(\Phi_g \Phi_g^T) \quad (\text{A.10})$$

where $\Phi_f \equiv \Theta_f^* - \Theta_f$ and $\Phi_g \equiv \Theta_g^* - \Theta_g$. Subtracting (40) from (27) and using (38) and (39), we obtain the error dynamic equation

$$\dot{\underline{e}} = -\alpha \underline{e} + \Phi_f \underline{p}(\underline{x}) + \Phi_g \underline{p}(\underline{x}) u + \underline{w}. \quad (\text{A.11})$$

The time derivative of V along the solution of (A.11) is

$$\begin{aligned} \dot{V} &= -\alpha |\underline{e}|^2 + \underline{e}^T \Phi_f \underline{p}(\underline{x}) + \underline{e}^T \Phi_g \underline{p}(\underline{x}) u + \underline{e}^T \underline{w} \\ &\quad + \frac{1}{\gamma_1} \text{tr}(\dot{\Phi}_f \Phi_f^T) + \frac{1}{\gamma_2} \text{tr}(\dot{\Phi}_g \Phi_g^T) \\ &= -\alpha |\underline{e}|^2 + \underline{e}^T \underline{w} + \frac{1}{\gamma_1} \text{tr}[(\gamma_1 \underline{e} \underline{p}^T(\underline{x}) - \dot{\Theta}_f) \Phi_f^T] \\ &\quad + \frac{1}{\gamma_2} \text{tr}[(\gamma_2 \underline{e} \underline{p}^T(\underline{x}) u - \dot{\Theta}_g) \Phi_g^T] \\ &= -\alpha |\underline{e}|^2 + \underline{e}^T \underline{w} + I_1^* \frac{\underline{e}^T \Theta_f \underline{p}(\underline{x})}{\text{tr}(\Theta_f \Theta_f^T)} \text{tr}(\Theta_f \Phi_f^T) \\ &\quad + I_2^* \frac{\underline{e}^T \Theta_g \underline{p}(\underline{x}) u}{\text{tr}(\Theta_g \Theta_g^T)} \text{tr}(\Theta_g \Phi_g^T) \end{aligned} \quad (\text{A.12})$$

where $I_1^* = 1$ ($I_2^* = 1$) if the second line of (41) [(42)] is true, $I_1^* = 0$ ($I_2^* = 0$) if the first line of (41) [(42)] is true, and we use the fact that $\dot{\Phi}_f = -\dot{\Theta}_f$. We now prove that $I_1^* (\underline{e}^T \Theta_f \underline{p}(\underline{x}) / \text{tr}(\Theta_f \Theta_f^T)) \text{tr}(\Theta_f \Phi_f^T) \leq 0$. If $I_1^* = 0$, the conclusion is trivial. For $I_1^* = 1$, which means that $\text{tr}(\Theta_f \Theta_f^T) = M_f$ and $\underline{e}^T \Theta_f \underline{p}(\underline{x}) > 0$, we have

$$\begin{aligned} \text{tr}(\Theta_f \Phi_f^T) &= \text{tr} \left[\Theta_f^* \Phi_f^T - \frac{1}{2} \Phi_f \Phi_f^T - \frac{1}{2} \Phi_f \Phi_f^T \right] \\ &= \frac{1}{2} \text{tr}(\Theta_f^* \Theta_f^T) - \frac{1}{2} \text{tr}(\Theta_f \Theta_f^T) - \frac{1}{2} \text{tr}(\Phi_f \Phi_f^T) \leq 0 \end{aligned} \quad (\text{A.13})$$

because $\text{tr}(\Theta_f^* \Theta_f^T) \leq M_f = \text{tr}(\Theta_f \Theta_f^T)$ and $\text{tr}(\Phi_f \Phi_f^T) \geq 0$. Hence, $I_1^* (\underline{e}^T \Theta_f \underline{p}(\underline{x}) / \text{tr}(\Theta_f \Theta_f^T)) \text{tr}(\Theta_f \Phi_f^T) \leq 0$. Similarly,

we can prove that $I_2^* (\underline{e}^T \Theta_g \underline{p}(\underline{x}) u / \text{tr}(\Theta_g \Theta_g^T)) \text{tr}(\Theta_g \Phi_g^T) \leq 0$. Therefore, from (A.12) we have

$$\dot{V} \leq -\alpha |\underline{e}|^2 + \underline{e}^T \underline{w} \leq -\alpha |\underline{e}|^2 + w_0 |\underline{e}| \quad (\text{A.14})$$

(recall that $w_0 \equiv \sup_{t \geq 0} |\underline{w}(t)|$ which is assumed to be finite). Hence, if $|\underline{e}| > (w_0/\alpha)$, we have $\dot{V} < 0$, which means that $\underline{e} \in L_\infty$. Since $\underline{e} = \underline{x} - \underline{\hat{x}}$ and \underline{x} is bounded by assumption, we have $\underline{\hat{x}} \in L_\infty$.

b) From (A.14) we have

$$\begin{aligned} \dot{V} &\leq -\frac{\alpha}{2} |\underline{e}|^2 - \frac{\alpha}{2} \left[|\underline{e}|^2 - \frac{2}{\alpha} \underline{e}^T \underline{w} + \frac{1}{\alpha^2} |\underline{w}|^2 - \frac{1}{\alpha^2} |\underline{w}|^2 \right] \\ &\leq -\frac{\alpha}{2} |\underline{e}|^2 + \frac{1}{2\alpha} |\underline{w}|^2. \end{aligned} \quad (\text{A.15})$$

Integrating both sides of (A.15), we have

$$\int_0^t |\underline{e}(\tau)|^2 d\tau \leq \frac{2}{\alpha} (|V(0)| + |V(t)|) + \frac{1}{\alpha^2} \int_0^t |\underline{w}(\tau)|^2 d\tau. \quad (\text{A.16})$$

Define $k_1 \equiv \frac{1}{\alpha} (|V(0)| + \sup_{t \geq 0} |V(t)|)$ and $k_2 \equiv 1/\alpha^2$, (A.16) becomes (45). k_1 is finite because \underline{e} , $\text{tr}(\Phi_f \Phi_f^T)$ and $\text{tr}(\Phi_g \Phi_g^T)$ are bounded from a), therefore $V \in L_\infty$ [see (A.10)].

c) If $\underline{w} \in L_2$, then from (45) we have that $\underline{e} \in L_2$. Because \underline{e} , Φ_f , Φ_g , $\underline{p}(\underline{x})$, u , and \underline{w} are all bounded ($\underline{p}(\underline{x})$ is bounded because from (37) we see that each element of $\underline{p}(\underline{x})$ is not greater than one), from (A.11) we have that $\dot{\underline{e}} \in L_\infty$. Using the Barbalat's Lemma [29] (if $\underline{e} \in L_2 \cap L_\infty$ and $\dot{\underline{e}} \in L_\infty$, then $\lim_{t \rightarrow \infty} |\underline{e}(t)| = 0$), we have $\lim_{t \rightarrow \infty} |\underline{e}(t)| = 0$. Q.E.D.

Proof of Theorem 4: a) From (60) and (62) we see that if $|\underline{\theta}_f|^2 = M_f$, then: if $\underline{e}^T ((\partial \hat{f} / \partial \underline{\theta}_f))^T \underline{\theta}_f \leq 0$, we have that $(d/dt)(|\underline{\theta}_f|^2) = 2 \underline{\theta}_f^T \dot{\underline{\theta}}_f = 2\gamma_1 \underline{e}^T ((\partial \hat{f} / \partial \underline{\theta}_f))^T \underline{\theta}_f \leq 0$; and, if $\underline{e}^T ((\partial \hat{f} / \partial \underline{\theta}_f))^T \underline{\theta}_f > 0$, we have that $(d/dt)(|\underline{\theta}_f|^2) = 2(\gamma_1 \underline{e}^T ((\partial \hat{f} / \partial \underline{\theta}_f))^T \underline{\theta}_f - \gamma_1 \underline{e}^T ((\partial \hat{f} / \partial \underline{\theta}_f))^T \underline{\theta}_f / |\underline{\theta}_f|^2 \underline{\theta}_f^T \underline{\theta}_f) = 0$. Hence, $|\underline{\theta}_f|^2 \leq M_f$. Using the same procedure we can prove that $|\underline{\theta}_g|^2 \leq M_g$. To prove $\underline{x} \in L_\infty$, define the Lyapunov function candidate

$$V_1 = \frac{1}{2} |\underline{e}|^2 + \frac{1}{2\gamma_1} |\underline{\phi}_f|^2 + \frac{1}{2\gamma_2} |\underline{\phi}_g|^2 \quad (\text{A.17})$$

where $\underline{\phi}_f \equiv \underline{\theta}_f^* - \underline{\theta}_f$ and $\underline{\phi}_g \equiv \underline{\theta}_g^* - \underline{\theta}_g$. Using (27) and (31) (replace Θ_f^* and Θ_g^* by $\underline{\theta}_f^*$ and $\underline{\theta}_g^*$), we obtain the error dynamic equation

$$\begin{aligned} \dot{\underline{e}} &= -\alpha \underline{e} + [\hat{f}(\underline{x}|\underline{\theta}_f^*) - \hat{f}(\underline{x}|\underline{\theta}_f)] \\ &\quad + [\hat{g}(\underline{x}|\underline{\theta}_g^*) - \hat{g}(\underline{x}|\underline{\theta}_g)] u + \underline{w}. \end{aligned} \quad (\text{A.18})$$

Using the Taylor series expansions of $\hat{f}(\underline{x}|\underline{\theta}_f^*)$ and $\hat{g}(\underline{x}|\underline{\theta}_g^*)$ around $\underline{\theta}_f$ and $\underline{\theta}_g$, we have

$$\hat{f}(\underline{x}|\underline{\theta}_f^*) - \hat{f}(\underline{x}|\underline{\theta}_f) = \left(\frac{\partial \hat{f}(\underline{x}|\underline{\theta}_f)}{\partial \underline{\theta}_f} \right)^T \underline{\phi}_f + \hat{f}_0, \quad (\text{A.19})$$

$$\hat{g}(\underline{x}|\underline{\theta}_g^*) - \hat{g}(\underline{x}|\underline{\theta}_g) = \left(\frac{\partial \hat{g}(\underline{x}|\underline{\theta}_g)}{\partial \underline{\theta}_g} \right)^T \underline{\phi}_g + \hat{g}_0 \quad (\text{A.20})$$

where \hat{f}_0 and \hat{g}_0 are the higher order terms. Using the Mean Value Theorem, we have

$$\begin{aligned}\hat{f}_0 &\equiv \hat{f}(x|\underline{\theta}_f^*) - \hat{f}(x|\underline{\theta}_f) - \left(\frac{\partial \hat{f}(x|\underline{\theta}_f)}{\partial \underline{\theta}_f} \right)^T \underline{\phi}_f \\ &= \left[\frac{\partial \hat{f}(x|\underline{\theta}_{f0})}{\partial \underline{\theta}_f} - \frac{\partial \hat{f}(x|\underline{\theta}_f)}{\partial \underline{\theta}_f} \right]^T \underline{\phi}_f, \quad (\text{A.21})\end{aligned}$$

$$\begin{aligned}\hat{g}_0 &\equiv \hat{g}(x|\underline{\theta}_g^*) - \hat{g}(x|\underline{\theta}_g) - \left(\frac{\partial \hat{g}(x|\underline{\theta}_g)}{\partial \underline{\theta}_g} \right)^T \underline{\phi}_g \\ &= \left[\frac{\partial \hat{g}(x|\underline{\theta}_{g0})}{\partial \underline{\theta}_g} - \frac{\partial \hat{g}(x|\underline{\theta}_g)}{\partial \underline{\theta}_g} \right]^T \underline{\phi}_g \quad (\text{A.22})\end{aligned}$$

where $\underline{\theta}_{f0} = \lambda_1 \underline{\theta}_f + (1 - \lambda_1) \underline{\theta}_f^*$ and $\underline{\theta}_{g0} = \lambda_2 \underline{\theta}_g + (1 - \lambda_2) \underline{\theta}_g^*$ for some $\lambda_1, \lambda_2 \in [0, 1]$. Substituting (A.19) and (A.20) into (A.18), we have

$$\begin{aligned}\dot{\underline{e}} &= -\alpha \underline{e} + \left(\frac{\partial \hat{f}(x|\underline{\theta}_f)}{\partial \underline{\theta}_f} \right)^T \underline{\phi}_f + \left(\frac{\partial \hat{g}(x|\underline{\theta}_g)}{\partial \underline{\theta}_g} \right)^T \\ &\quad \cdot \underline{\phi}_g u + \hat{f}_0 + \hat{g}_0 u + \underline{w}. \quad (\text{A.23})\end{aligned}$$

The time derivative of V_1 along the solution of (A.23) is

$$\begin{aligned}\dot{V}_1 &= -\alpha |\underline{e}|^2 + \underline{e}^T \left(\frac{\partial \hat{f}}{\partial \underline{\theta}_f} \right)^T \underline{\phi}_f + \underline{e}^T \left(\frac{\partial \hat{g}}{\partial \underline{\theta}_g} \right)^T \underline{\phi}_g u + \underline{e}^T \hat{f}_0 \\ &\quad + \underline{e}^T \hat{g}_0 u + \underline{e}^T \underline{w} \\ &\quad + \frac{1}{\gamma_1} \dot{\underline{\phi}}_f^T \underline{\phi}_f + \frac{1}{\gamma_2} \dot{\underline{\phi}}_g^T \underline{\phi}_g. \quad (\text{A.24})\end{aligned}$$

Using (60)–(63) and $\dot{\underline{\phi}}_f = -\dot{\underline{\theta}}_f$, $\dot{\underline{\phi}}_g = -\dot{\underline{\theta}}_g$, we have that

$$\begin{aligned}\dot{V}_1 &= -\alpha |\underline{e}|^2 + \underline{e}^T \hat{f}_0 + \underline{e}^T \hat{g}_0 u + \underline{e}^T \underline{w} + I_1^* \frac{\underline{e}^T \left(\frac{\partial \hat{f}}{\partial \underline{\theta}_f} \right)^T \underline{\theta}_f}{|\underline{\theta}_f|^2} \underline{\theta}_f^T \underline{\phi}_f \\ &\quad + I_2^* \frac{\underline{e}^T \left(\frac{\partial \hat{g}}{\partial \underline{\theta}_g} \right)^T \underline{\theta}_g u}{|\underline{\theta}_g|^2} \underline{\theta}_g^T \underline{\phi}_g \quad (\text{A.25})\end{aligned}$$

where $I_1^* = 0$ (1) if the condition of the first (second) line of (60) is true, and $I_2^* = 0$ (1) if the condition of the first (second) line of (61) is true. Using the same arguments as in the proof of Theorem 3 and noticing that $|\underline{\theta}|^2 = \text{tr}(\underline{\theta} \underline{\theta}^T)$, we can prove that the last two terms of (A.25) are nonpositive. Hence

$$\begin{aligned}\dot{V}_1 &\leq -\alpha |\underline{e}|^2 + \underline{e}^T \hat{f}_0 + \underline{e}^T \hat{g}_0 u + \underline{e}^T \underline{w} \\ &\leq -\alpha |\underline{e}|^2 + |\underline{e}| \cdot |\hat{f}_0| + |\underline{e}| \cdot |\hat{g}_0| \cdot |u| + |\underline{e}| \cdot |\underline{w}|. \quad (\text{A.26})\end{aligned}$$

From (54)–(59) we see that all the elements of $(\partial \hat{f} / \partial \underline{\theta})$ and $(\partial \hat{g} / \partial \underline{\theta})$ are bounded; therefore, from (A.21) and (A.22) and the boundedness of $\underline{\phi}_f$ and $\underline{\phi}_g$ (we have proven that $|\underline{\theta}_f|^2 \leq M_f$ and $|\underline{\theta}_g|^2 \leq M_g$; $|\underline{\theta}_f^*|^2 \leq M_f$ and $|\underline{\theta}_g^*|^2 \leq M_g$ by definition), we have that $|\hat{f}_0|$ and $|\hat{g}_0|$ are bounded. Hence, $\beta \equiv |\hat{f}_0| + |\hat{g}_0| \cdot |u| + |\underline{w}|$ is finite. From (A.26) we have that if $|\underline{e}| > \beta / \alpha$, then $\dot{V}_1 < 0$; therefore, $|\underline{e}|$ is bounded. Since $\underline{\hat{x}} = \underline{x} - \underline{e}$ and \underline{x} is bounded by assumption, we have $\underline{\hat{x}} \in L_\infty$.

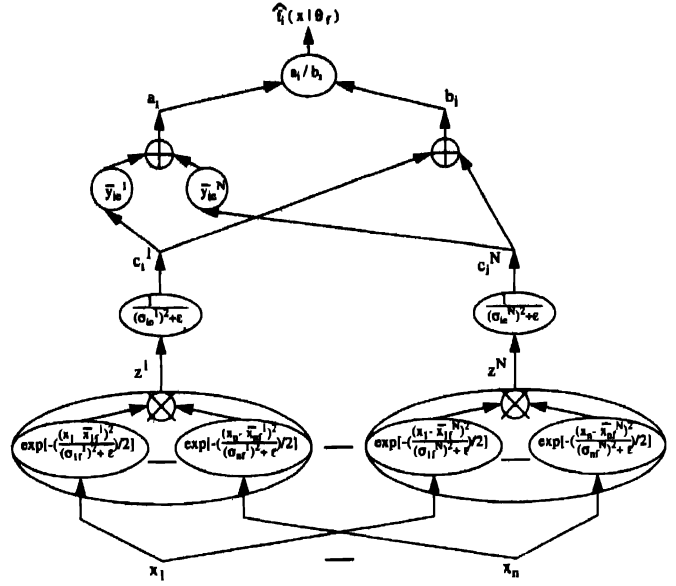


Fig. 13. Network representation of the fuzzy system.

b) From (A.21), (A.22), (65), and the first line of (A.26), we have that

$$\begin{aligned}\dot{V}_1 &\leq -\frac{\alpha}{2} |\underline{e}|^2 - \frac{\alpha}{2} \left(|\underline{e}|^2 - \frac{2}{\alpha} \underline{e}^T \underline{v} + \frac{1}{\alpha^2} |\underline{v}|^2 - \frac{1}{\alpha^2} |\underline{v}|^2 \right) \\ &\leq -\frac{\alpha}{2} |\underline{e}|^2 + \frac{1}{2\alpha} |\underline{v}|^2. \quad (\text{A.27})\end{aligned}$$

Integrating both sides of (A.27) and defining $k_1 \equiv (1/\alpha)(|V_1(0)| + \sup_{t>0} |V(t)|)$ and $k_2 \equiv 1/\alpha^2$, we obtain (64) ($\sup_{t>0} |V(t)|$ is finite because \underline{e} , $\underline{\phi}_f$ and $\underline{\phi}_g$ are all bounded).

c) If $\underline{v} \in L_2$, then from (64) $\underline{e} \in L_2$. From a), $\underline{e} \in L_\infty$. Because in the proof of b) of this theorem we showed that all the variables in the right-hand side of (A.23) are bounded, we have $\dot{\underline{e}} \in L_\infty$. Using the Barbalat's Lemma, we have $\lim_{t \rightarrow \infty} |\underline{e}(t)| = 0$. Q.E.D.

APPENDIX B

The fuzzy system $f_i(x|\underline{\theta}_f)$ of (52) can be represented as the feedforward network of Fig. 13. Using the chain rule on the \hat{f}_i of Fig. 13, we have that

$$\frac{\partial \hat{f}_i}{\partial \bar{y}_{ie}^l} = \frac{c_i^l}{b_i}, \quad (\text{B.1})$$

$$\frac{\partial \hat{f}_i}{\partial \sigma_{ie}^l} = \frac{\partial \hat{f}_i}{\partial c_i^l} \frac{\partial c_i^l}{\partial \sigma_{ie}^l} = \frac{(\bar{y}_{ie}^l - \hat{f}_i)}{b_i} \frac{-2z^l \sigma_{ie}^l}{((\sigma_{ie}^l)^2 + \epsilon)}, \quad (\text{B.2})$$

$$\begin{aligned}\frac{\partial \hat{f}_i}{\partial \bar{x}_{jf}^l} &= \frac{\partial \hat{f}_i}{\partial z^l} \frac{\partial z^l}{\partial \bar{x}_{jf}^l} = \frac{\bar{y}_{ie}^l - \hat{f}_i}{b_i} \frac{1}{(\sigma_{ie}^l)^2 + \epsilon} \\ &\quad \cdot \frac{z^l (x_j - \bar{x}_{jf}^l)}{(\sigma_{jf}^l)^2 + \epsilon}, \quad (\text{B.3})\end{aligned}$$

$$\begin{aligned}\frac{\partial \hat{f}_i}{\partial \sigma_{jf}^l} &= \frac{\partial \hat{f}_i}{\partial z^l} \frac{\partial z^l}{\partial \sigma_{jf}^l} = \frac{\bar{y}_{ie}^l - \hat{f}_i}{b_i} \frac{1}{(\sigma_{ie}^l)^2 + \epsilon} \\ &\quad \cdot \frac{z^l 2(x_j - \bar{x}_{jf}^l) \sigma_{jf}^l}{((\sigma_{jf}^l)^2 + \epsilon)^2} \quad (\text{B.4})\end{aligned}$$

where $j = 1, 2, \dots, n$, $l = 1, 2, \dots, N$, and a_i , b_i , c_i^l , and z^l are defined in Fig. 13. The $(\partial g_i / \partial \theta_g)$ can be computed in exactly the same way as (B.1)–(B.4), with the subscripts “e” replaced by “h” and “f” replaced by “g.”

REFERENCES

- [1] R. E. Bellman and L. A. Zadeh, “Local and fuzzy logics,” in *Modern Uses of Multiple-Valued Logic*, J. M. Dunn and G. Epstein, Eds., Dordrecht, Netherlands: Reidel, 1977, pp. 103–165.
- [2] J. J. Buckley, “Universal fuzzy controllers,” *Automatica*, vol. 28, no. 6, pp. 1245–1248, 1992.
- [3] S. Chen, C. F. N. Cowan, and P. M. Grant, “Orthogonal least squares learning algorithm for radial basis function networks,” *IEEE Trans. Neural Networks*, vol. 2, no. 2, pp. 302–309, 1991.
- [4] Y. Y. Chen, “The global analysis of fuzzy dynamic systems,” Ph.D. dissertation, Dept. Elec. Eng., Univ. California, Berkeley, 1989.
- [5] S. Chiu, S. Chand, D. Moore, and A. Chaudhary, “Fuzzy logic for control of roll and moment for a flexible wing aircraft,” *IEEE Contr. Syst. Mag.*, vol. 11, no. 4, pp. 42–48, 1991.
- [6] D. Dubois and H. Prade, *Fuzzy Sets and Systems: Theory and Applications*. Orlando, FL: Academic, 1980.
- [7] G. C. Goodwin and D. Q. Mayne, “A parameter estimation perspective of continuous time model reference adaptive control,” *Automatica*, vol. 23, pp. 57–70, 1987.
- [8] G. C. Goodwin and R. L. Payne, *Dynamic System Identification: Experiment Design and Data Analysis*. New York: Academic, 1977.
- [9] J. K. Hale, *Ordinary Differential Equations*. New York: Wiley-Interscience, 1969.
- [10] A. V. Holden, *Chaos*. Princeton, NJ: Princeton Univ. Press, 1986.
- [11] C. Isik, “Identification and fuzzy rule-based control of a mobile robot motion,” in *Proc. IEEE Int. Symp. Intelligent Control*, Philadelphia, PA, 1987.
- [12] W. Kickert and E. Mamdani, “Analysis of a fuzzy logic controller,” *Fuzzy Sets and Syst.*, vol. 1, no. 1, pp. 29–44, 1978.
- [13] M. Kinoshita, T. Fukuzaki, T. Satoh, and M. Miyake, “An automatic operation method for control rods in BWR plants,” in *Proc. Specialists’ Meeting on In-Core Instrumentation and Reactor Core Assessment*, Cadarache, France, 1988.
- [14] J. Kiszka, M. Gupta, and P. Nikiforuk, “Energetic stability of fuzzy dynamic systems,” *IEEE Trans. Syst., Man, Cybernetics*, vol. SMC-15, no. 5, pp. 783–792, 1985.
- [15] G. J. Klir and T. A. Folger, *Fuzzy Sets, Uncertainty, and Information*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
- [16] G. Langari and M. Tomizuka, “Stability of fuzzy linguistic control systems,” in *Proc. IEEE Conf. Decis. Contr.*, Hawaii, 1990, pp. 2185–2190.
- [17] P. M. Larsen, “Industrial application of fuzzy logic control,” *Int. J. Man Mach. Studies*, vol. 12, no. 1, pp. 3–10, 1980.
- [18] C. C. Lee, “Fuzzy logic in control systems: Fuzzy logic controller, part I,” *IEEE Trans. Syst., Man, Cybernetics*, vol. 20, no. 2, pp. 404–418, 1990.
- [19] ———, “Fuzzy logic in control systems: Fuzzy logic controller, part II,” *IEEE Trans. Syst., Man, Cybernetics*, vol. 20, no. 2, pp. 419–435, 1990.
- [20] L. Ljung, *System Identification—Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
- [21] E. H. Mamdani, “Applications of fuzzy algorithms for simple dynamic plant,” in *Proc. IEE*, vol. 121, no. 12, 1974, pp. 1585–1588.
- [22] E. H. Mamdani, “Application of fuzzy logic to approximate reasoning using linguistic systems,” *IEEE Trans. Computers*, vol. 26, pp. 1182–1191, 1977.
- [23] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [24] K. S. Narendra and K. Parthasarathy, “Identification and control of dynamical systems using neural networks,” *IEEE Trans. Neural Networks*, vol. 1, no. 1, pp. 4–27, 1990.
- [25] C. V. Negoita and P. A. Ralescu, *Application of Fuzzy Sets to System Analysis*. ISR. II. Basel: Birkhauser, 1975.
- [26] M. M. Polycarpou and P. A. Ioannou, “Identification and control nonlinear systems using neural network models: Design and stability analysis,” EE-Report 91-09-01, Univ. Southern California, 1991.
- [27] M. J. D. Powell, “Radial basis functions for multivariable interpolation: A review,” in *Algorithms for Approximation*, J. C. Mason and M. G. Cox, Eds., New York: Oxford, 1987, pp. 143–167.
- [28] R. M. Sanner and J. E. Slotine, “Gaussian networks for direct adaptive control,” in *Proc. Amer. Contr. Conf.*, 1991, pp. 2153–2159.
- [29] S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence, and Robustness*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [30] H. Takagi, “Introduction to Japanese consumer products that apply neural networks and fuzzy systems,” personal communication, 1992.
- [31] H. Takagi and M. Sugeno, “Fuzzy identification of systems and its applications to modeling and control,” *IEEE Trans. Syst., Man, Cybern.*, vol. 15, no. 1, pp. 116–132, 1985.
- [32] R. M. Tong, “A control engineering review of fuzzy systems,” *Automatica*, vol. 13, pp. 559–569, 1977.
- [33] ———, “Some properties of fuzzy feedback systems,” *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-10, no. 6, pp. 327–330, 1980.
- [34] L. X. Wang, *Adaptive Fuzzy Systems and Control: Design and Stability Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1994.
- [35] ———, “Fuzzy systems are universal approximators,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, San Diego, 1992, pp. 1163–1170.
- [36] ———, “Stable adaptive fuzzy control of nonlinear systems,” *IEEE Trans. Neural Networks*, vol. 1, no. 2, pp. 146–155, 1993.
- [37] L. X. Wang and J. M. Mendel, “Generating fuzzy rules by learning from examples,” *IEEE Trans. Syst., Man, Cybern.*, vol. 22, no. 6, pp. 1414–1427, 1992.
- [38] ———, “Back-propagation fuzzy systems as nonlinear dynamic system identifiers,” in *Proc. IEEE Int. Conf. Fuzzy Syst.*, San Diego, 1992, pp. 1409–1418.
- [39] ———, “Fuzzy basis function, universal approximation, and orthogonal least squares learning,” *IEEE Trans. Neural Networks*, vol. 3, no. 5, pp. 807–814, 1992.
- [40] S. Yasunobu, S. Miyamoto, and H. Ihara, “Fuzzy control for automatic train operation system,” in *Proc. 4th IFAC/IFIP/IFORS Int. Congress on Control in Transportation Systems*, Baden-Baden, 1983.
- [41] L. A. Zadeh, “Fuzzy sets and systems,” in *Proc. Symp. Syst. Theory*, Polytech. Inst. Brooklyn, 1965, pp. 29–37.
- [42] ———, “Fuzzy sets,” *Informat. Contr.*, vol. 8, pp. 338–353, 1965.
- [43] ———, “A rationale for fuzzy control,” *J. Dyn. Syst. Meas. Contr.*, vol. 34, pp. 3–4, 1972.
- [44] ———, “Outline of a new approach to the analysis of complex systems and decision processes,” *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 1, pp. 28–44, 1973.
- [45] ———, “A theory of approximating reasoning,” in *Machine Intelligence*, J. E. Hayes, D. Michie, and L. I. Mikulich, Eds., vol. 9. New York: Elsevier, 1979, pp. 149–194.
- [46] X. J. Zeng and M. G. Singh, “Approximation theory of fuzzy systems—MIMO case,” personal communication, 1993.

Li-Xin Wang (S'90–M'93) received the B.S. and M.S. degrees from Northwestern Polytechnical University, Xian, People's Republic of China, in 1984 and 1987, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, in 1992, all in electrical engineering.

From 1987 to 1989 he was with the Department of Computer Science and Engineering, Northwestern Polytechnical University. From the fall of 1989 to the spring of 1992 he was a Research/Teaching Assistant with the Department of Electrical Engineering–Systems, University of Southern California, where he was working toward the Ph.D. degree. From the summer of 1992 to the summer of 1993 he was a postdoctoral fellow in the Department of Electrical Engineering and Computer Science, University of California at Berkeley, where he worked with Prof. Zadeh. Since the fall of 1993 he has been an Assistant Professor with the Department of Electrical and Electronic Engineering, the Hong Kong University of Science and Technology, Hong Kong. His research interests include intelligent control, fuzzy systems, and robotics.

Dr. Wang is author of *Adaptive Fuzzy Systems and Control: Design and Stability Analysis* (Prentice-Hall, 1994). He received a Phi Kappa Phi Student Recognition Award in 1992 for his work on fuzzy systems.

Stochastic Approximation with Averaging and Feedback: Rapidly Convergent "On-Line" Algorithms

Harold J. Kushner, *Fellow, IEEE*, and Jichuan Yang

Abstract— Consider the stochastic approximation $X_{n+1} = X_n + a_n g(X_n, \xi_n)$, where $0 < a_n \rightarrow 0$, $\sum_n a_n = \infty$ and $\{\xi_n\}$ is the "noise" sequence. Suppose that $a_n \rightarrow 0$ slowly enough such that $a_n/a_{n+1} = 1 + o(a_n)$. Then various authors have shown that the rate of convergence of the average $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ is optimal in the sense that $\sqrt{n}(\bar{X}_n - \theta)$ converged in distribution to a normal random variable with mean zero and covariance V , where V was the smallest possible in an appropriate sense. V did not depend on $\{a_n\}$. The analogs of the advantages of averaging extend to the constant parameter systems $X_{n+1} = X_n + \epsilon g(X_n, \xi_n)$ for small $\epsilon > 0$. The averaging method is essentially "off line" in the sense that the actual SA iterate X_n is not influenced by the averaging. In many applications, X_n itself is of greatest interest, since that is the "operating parameter." This paper deals with the problem of stochastic approximation with averaging and with appropriate feedback of the averages into the original algorithm. It is shown both mathematically and via simulation that it works very well and has numerous advantages. It is a clear improvement over the system X_n by itself. It is fairly robust, and quite often it is much preferable to the use of the above averages without feedback. We will deal, in particular, with "linear" algorithms of the type appearing in parameter estimators, adaptive noise cancellers, channel equalizers, adaptive control, and similar applications. The main development will be for the constant parameter case because of its importance in applications. But analogous results hold for the case where $a_n \rightarrow 0$.

1. INTRODUCTION

CONSIDER the stochastic approximation

$$X_{n+1} = X_n + a_n g(X_n, \xi_n)$$

where $0 < a_n \rightarrow 0$, $\sum_n a_n = \infty$ and $\{\xi_n\}$ is the "noise" sequence. There is a large literature and many sophisticated analyses available for such processes. Under various sets of conditions, there is θ such that $X_n \rightarrow \theta$ either with probability one or weakly. Also, under appropriate conditions $(X_n - \theta)/\sqrt{a_n}$ converges in distribution to a normal random variable with mean zero and some covariance matrix V_0 . The matrix V_0 and the scale factors $\{a_n\}$ are a measure of the "rate of convergence."

Suppose that $a_n \rightarrow 0$ slowly relative to $1/n$. More particularly, let

$$a_n/a_{n+1} = 1 + o(a_n). \quad (1.2)$$

Manuscript received December 3, 1992; revised January 15, 1994. Recommended by Past Associate Editor, W. S. Wong. This work was supported in part by AFOSR Contract F49620-92-0081 and NSF Grant ECS-8913351.

The authors are with the Division of Applied Mathematics, Brown University, Providence, RI 02912 USA.
IEEE Log Number 9406096.

Under some additional conditions, Polyak and Juditsky [11] showed that the rate of convergence of the average

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (1.3)$$

is optimal in an important sense. In particular, $\sqrt{n}(\bar{X}_n - \theta)$ converged in distribution to a normal random variable with mean zero and covariance V , where V was the smallest possible in an appropriate sense. V did not depend on $\{a_n\}$. This development promises to be quite important in applications. By now, extensive simulations have supported the theoretical conclusions. This superiority would not be true if a_n decreased as $O(1/n)$. Choosing the a_n sequence was a major nuisance in the past. The importance of this problem is now much reduced. Larger values of a_n (larger than $O(1/n)$) cause the process to "jump around" more, but the averaging in (1.3) compensates for this and greatly improves the values. Ruppert [12] developed similar results for a one dimensional problem.

Additional work on the averaging problem was done by Yin [14], [15] who generalized the earlier results. In [9], Kushner and Yang showed that the averaging results were generic to stochastic approximation (SA), and held under conditions of considerable generality. That work shed much light on the reasons why the averaging gave a better result when (1.2) held. In general, that reference used sums of the type (1.3), but with various moving windows and where the lower index of summation goes to infinity as $n \rightarrow \infty$. As pointed out in [9], the basic reasons for the success of averaging stem from the fact that the "time scales" of the original sequence X_n and the averaged sequence \bar{X}_n can be separated, with the time scale of the former sequence being "faster." This time scale separation also accounts for the difficulties in the analysis of the averaged sequence. Time scale separations have been used as a standard tool in the analysis of stochastic approximations. But, in the past it was the separation of the scales of the X_n and the noise ξ_n sequences which was used. Now there is a third time scale to be considered. Similar results have been obtained for the constant parameter case

$$X_{n+1} = X_n + \epsilon g(X_n, \xi_n). \quad (1.4)$$

The averaging method is essentially "off line" in the sense that the iterations (1.1) and (1.4) are not influenced by the averaging. In many applications, X_n itself is of greatest interest since that is the "operating parameter." Indeed, if \bar{X}_n were substituted for X_n on the right sides of (1.1) or

(1.4), or in the $g(\cdot)$ function alone, then the advantages of averaging would be totally lost. One can, of course, run a "training" sequence and then use the end value of the averaged sequence as the future operating value. This would often be a useful idea. Nevertheless, the question remains whether it is possible to exploit the averaging to improve the actual primary stochastic approximation $\{X_n\}$. This paper deals with the problem of stochastic approximation with simultaneous use of averaging and feedback. It is shown both mathematically and via simulation that it works well and has numerous advantages. It is a clear improvement over the systems (1.1), (1.4) by themselves.

We will deal only with "linear" algorithms of the type appearing in parameter estimators, adaptive noise cancellers, adaptive antenna arrays, channel equalizers, and similar applications. Stability problems make the general nonlinear problem much harder at present. But, as noted above, the special case has numerous important applications and is the subject of an enormous literature itself. These are adaptive problems of great current interest, particularly in the communications, parameter estimation and signal processing area. For such problems, it is common practice to use (1.4) for small ϵ , even if the parameters of concern are not time varying. So, we concentrate on this case. More detailed discussion and motivation for the constant parameter case is given at the end of Section III to which the reader is referred.

The multiple time scale point of view and the clear (to be seen below) advantages of feedback with which this paper is concerned imply that the use of various types of averaging methods in stochastic approximation is just beginning. In Section II we describe a simple but important averaging result (with no feedback), taken from [9]. While this result will not be used explicitly in the sequel, it illustrates a basic property and value of averaging in a simple context and it is useful for motivation. In Section III the actual problem of interest will be described. We work with (1.4) and use various approximations to (1.3) which are equivalent to different windows of averaging or "forgetting rates." All the results have analogs for the case where $a_n \rightarrow 0$ and (1.2) holds. The basic estimates of the asymptotic orders of the errors for the problem with feedback is given in Section IV. This section makes use of the so-called perturbed Lyapunov function methods. The order estimates show that the averaged iterate (in the feedback case) is still better than what one would get without averaging. Once the estimates of order are available, an asymptotic weak convergence analysis can be done. This appears in Section V, where the limits of the distributions of normalized iterates are obtained. In particular, it will be seen that the limiting distribution is that of the stationary random variable of a Gauss-Markov diffusion, and that the limit covariance of the process with feedback is much better than that for the original sequence without feedback and frequently better than that for the sequence of averages (1.3) of the original sequence (1.1). Numerous simulations have strongly supported the theoretical conclusions, and typical simulation results for canonical systems fitting the framework of the various applications cited above appear in Section VI. The time varying parameter case is discussed in Section V-B.

II. THE ADVANTAGES OF AVERAGING: A SIMPLE EXAMPLE

A very useful and relatively simple result which demonstrates the advantages of averaging without feedback can be readily obtained via a weak convergence method, under quite broad conditions. We will use a "minimal window" of iterates, smaller than in (1.3), yet of considerable practical interest. Our interest in this section is simply in showing that various sets of conditions in stochastic approximation give the desired result. We wish to show only that the averaging works if there is a basic rate of convergence result for X_n , thus implying that the advantages of averaging naturally accompany any classical rate of convergence analysis. The material in this section is taken from [9] and is for motivation only. The result will not be used explicitly in the sequel, but is important for motivation and for putting the subsequent results into perspective.

Let us set up the required terminology. Define the "interpolated time" $t_n = \sum_{i=0}^{n-1} a_i$, with $t_0 = 0$. In this section, we will set $\theta = 0$ [see the definition below (1.1)] for notational simplicity. In the asymptotic analysis, it is usual to examine the tails of the SA sequences. To do this we define the "tail" processes as follows. For each $n \geq 0$, define the interpolated processes $U^n(\cdot)$, $X^n(\cdot)$ by

$$\begin{aligned} U^n(t) &= X_{n+i} / \sqrt{a_{n+i}} \text{ for} \\ t &\in [t_{n+i} - t_n, t_{n+i+1} - t_n], i \geq 0, \\ X^n(t) &= X_{n+i} \text{ for } t \in [t_{n+i} - t_n, t_{n+i+1} - t_n], i \geq 0. \end{aligned}$$

The basic result below is stated in terms of weak convergence. It is not necessary to be familiar with the details of the weak convergence method to deal with the rest of the paper. Loosely speaking, weak convergence of a sequence of processes $U^n(\cdot)$ to a process $U(\cdot)$ means that if $f(\cdot)$ is a continuous functional of the process paths, then $f(U^n(\cdot))$ converges in distribution to $f(U(\cdot))$. It is a considerable extension of the usual convergence in distribution for vector valued random variables. Of course, to speak of convergence in a function or path space implies a particular path space and topology on it. The most frequently used path space is the space of functions which are continuous on the right and with left hand limits. The most commonly used topology on this space (and the one used here) is the so-called Skorohod topology $D^r[0, \infty)$ [2], [3], [6], where r is the dimension of X . All that we need to know here is that the topology is weaker than that determined by uniform convergence on bounded time intervals. Thus, if $y_n(\cdot)$ is a sequence of paths in $D^r[0, \infty)$ and $y_n(\cdot) \rightarrow y(\cdot)$ in the sup norm on each bounded time interval, then the convergence is also in the Skorohod topology. Let \Rightarrow denote weak convergence.

In Theorem 2.1, we will use the following assumption.

A 2.1: There is a matrix G whose eigenvalues lie in the open left-half plane and a positive semidefinite symmetric matrix R_0 such that $X^n(\cdot) \Rightarrow$ zero process and $U^n(\cdot) \Rightarrow U(\cdot)$, where $U(\cdot)$ is the stationary solution to

$$dU = GU dt + R_0^{\frac{1}{2}} dw \quad (2.1)$$

where $w(\cdot)$ is a standard R^r -valued Wiener process.

Comment on (A2.1): We state condition (A2.1) in the given form since it is the most convenient way to illustrate the main point. Many different sets of conditions imply (A2.1) and getting such a result has been one of the main goals of work in stochastic approximation. Our primary goal in this section is to show that such a commonly available weak convergence result can be easily extended to show the possibilities in improvement in the rate of convergence due to averaging. The references [1], [6], [7], [10] contain various sets of conditions which guarantee (A2.1).

For $T > 0$, define $Z^n(\cdot)$ by

$$Z^n(T) = \frac{1}{\sqrt{T/a_n}} \sum_{i=n}^{n+T/a_n} X_i. \quad (2.2)$$

In sums such as (2.2), the limits of summation are always taken to be the integer parts of the indexes.

A. A Convergence Theorem

Theorem 2.1: Assume (1.2) and (A2.1) and define $V = G^{-1}R_0(G')^{-1}$. For each T , $Z^n(T)$ converges in distribution to a random variable with mean zero and covariance $V_T = V + O(1/T)$.

Discussion: The proof of the theorem is in [9]. By the weak convergence in (A2.1), as $n \rightarrow \infty$ we have (loosely speaking and in the sense of distributions) $X_n \sim N(0, a_n V_0)$, where $N(0, M)$ denotes the normal distribution with covariance M and mean zero, and V_0 is the stationary covariance of (2.1). The theorem implies that for large T and large n , $\bar{X}_n \sim N(0, a_n V/T)$. Thus for large T , averaging is a distinct advantage. The sum in (2.2) can be replaced by $\sum_{i=n-T/a_n}^n$. Other relevant comments are in [9], [11], [14], with various windows of averaging.

Minimality of V : It is noted in [9], [11] that V is minimal in an important sense, which we now describe. Suppose that (1.1) is used and $U_n = X_n/\sqrt{a_n}$ converged in distribution to a normally distributed random variable \tilde{U} with mean zero. Then the fastest asymptotic rate of convergence is obtained with $a_n = K/n$, for K a positive definite matrix. Then, under appropriate conditions (see, e.g., [1], [6]) $U^n(\cdot) \Rightarrow \tilde{U}(\cdot)$, where $\tilde{U}(\cdot)$ is the stationary solution to

$$d\tilde{U} = \left(\frac{I}{2} + KG\right)\tilde{U}dt + KR_0^{\frac{1}{2}}dw \quad (2.3)$$

where G and R_0 are as in (A2.1), and it is supposed that $(\frac{I}{2} + KG)$ is a stable matrix. If we optimize the trace of the covariance matrix of (2.3) over K , we get the best value of K as

$$K = -G^{-1}.$$

With this value of K , the covariance of $\tilde{U}(0)$ is just the V used in Theorem 2.1. In this sense, the result in [9], [11] and of Theorem 2.1 is optimal.

Constant Gains: Suppose that $a_n = \varepsilon$, a small gain parameter. Define

$$Z^n(T) = \frac{1}{\sqrt{T/\varepsilon}} \sum_{i=n}^{n+T/\varepsilon} X_i.$$

Then if n and T are large enough, we have $X_n \sim N(0, \varepsilon V_0)$, and $\bar{X}_n \sim N(0, \varepsilon V/T)$, again an advantage over the nonaveraged variate.

III. AVERAGING WITH FEEDBACK FORMULATION

In this section, we describe the algorithm of concern. The analysis will be done in the following section. We would like to be concerned with algorithms of the general type, for positive real valued A

$$X_{n+1} = X_n + \varepsilon g(X_n, \xi_n) + \varepsilon A(\bar{X}_n - X_n)$$

where the feedback is linear and proportional to the gain ε . This form does allow the use of the averaged iterates in an intuitively reasonable way in the primary algorithm. Indeed, the results of the sequel suggest that a larger feedback will not work well. It is hard to analyze the fully nonlinear case, and we will work with a simpler form which is very important in applications.

The Model: Consider a classical algorithm for parameter identification. Let θ denote the system parameter, ϕ_n a sequence of random "input vectors" with $E\phi_n\phi_n' = G > 0$ (positive definite), ψ_n a sequence of real valued random variables (with mean zero, without loss of generality) and ϕ' is the transpose of ϕ . The observed output is the real valued $y_n \equiv \theta' \phi_n + \psi_n$. A common form of the identification (alternatively, adaptive noise canceller, etc.) algorithm is

$$\hat{\theta}_{n+1} = \theta_n + \varepsilon \phi_n [y_n - \hat{\theta}_n' \phi_n] \quad (3.1)$$

where $\hat{\theta}_n$ is the n th estimate of θ . Define

$$\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_i. \quad (3.2)$$

One also might use the finite window of averaging, where $\bar{\theta}_n$ is defined by (for $T > 0$)

$$\bar{\theta}_n = \frac{1}{T/\varepsilon} \sum_{i=n-T/\varepsilon}^n \hat{\theta}_i.$$

Then (see comment after the proof of Theorem 2.1) the theory of the last section (equivalently, of [9]) can be used to show that for large T , the average $\bar{\theta}_n$ has a much smaller asymptotic covariance than does the original $\hat{\theta}_n$.

A Feedback Form of (3.1), (3.2). Consider the following form of (3.1) and (3.2), where the average is fed back. For $A > 0$, redefine θ_n and $\bar{\theta}_n$ by

$$\begin{aligned} \hat{\theta}_{n+1} &= \hat{\theta}_n + \varepsilon \phi_n [y_n - \hat{\theta}_n' \phi_n] + \varepsilon A[\bar{\theta}_n - \theta_n] \\ \bar{\theta}_n &= \frac{1}{n} \sum_{i=1}^n \hat{\theta}_i. \end{aligned} \quad (3.3)$$

Defining $X_n = \hat{\theta}_n - \theta$, $\bar{X}_n = \bar{\theta}_n - \theta$, rewrite (3.3) as

$$X_{n+1} = [I - \varepsilon G]X_n + \varepsilon \xi_n + \varepsilon A[\bar{X}_n - X_n], \quad (3.4a)$$

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i. \quad (3.4b)$$

where

$$\xi_n = [G - \phi_n \phi_n'] X_n + \phi_n \psi_n \equiv \Phi_n X_n + \rho_n$$

where the definitions of Φ_n and ρ_n are obvious.

The sequence $\{X_n\}$ can be arbitrarily well approximated for large n by the solution to the equation

$$\begin{aligned} \bar{X}_0 &= \bar{X}_1 = X_0 \quad \text{and, for } n > 1, \\ \bar{X}_{n+1} &= \left(1 - \frac{1}{n}\right) \bar{X}_n + \frac{X_n}{n}. \end{aligned} \quad (3.5)$$

In fact, we will work with the more general forms defined by either (3.6) or (3.7)

$$\bar{X}_{n+1} = (1 - b_n) \bar{X}_n + b_n X_n \quad (3.6)$$

where

$$b_n \rightarrow 0, \quad \sum b_n = \infty$$

or

$$\bar{X}_{n+1} = (1 - \alpha c) \bar{X}_n + \alpha c X_n \quad (3.7)$$

where $\alpha > 0$ is small. The form

$$\bar{X}_n = \alpha c \sum_{i=0}^{n+1/\alpha c} X_i$$

seems to be better in simulations although it is much harder to analyze.

Algorithms of the form (3.1) are commonly used in adaptive applications throughout communication theory and in signal processing [13], [5]. The simulations in Section VI are for a canonical problem in these areas. In such applications, one generally lets ϵ be a constant, selected *a priori*. The decreasing gain case $a_n \rightarrow 0$ is more rarely used, even when the basic parameters to be adjusted to are thought to be constant. This is due partly to the way that the systems are automated. Also, in applications one often seeks the largest value of ϵ for which the system will be stable and still have reasonably small noise effects. It is important to keep in mind that the errors are relative to a bias, since the system order is rarely known. (In this biased case the results here hold, as they should, for the errors centered about the bias.)

A possibly ideal algorithm, in lieu of (3.1), would be linearized least squares. This has excellent asymptotic properties, but due to the heavy computational overhead, one generally prefers to avoid it in the applications cited. A possible alternative to linearized least squares replaces ϵ by $a_n \rightarrow 0$. This is often unsatisfactory. The best rate is $a_n = O(1/n)$, which has serious robustness problems. Furthermore, unless $a_n = K/n$ for an appropriate matrix K , the data for the different coordinates can be greatly skewed. In many applications one prefers not to use such a matrix due to the more complex implementation and the difficulty in getting a good value for it.

As seen from Section II, letting a_n go to zero slowly enough and using averaging accomplishes the same thing asymptotically as does the use of linearized least squares,

without much additional computation, since (loosely speaking) $X_n \sim N(0, V/n)$, where V is minimal. The best algorithms of Section II use a window of averaging which is of the order of n , the current iterate number. One might wish to avoid this due to robustness considerations. In addition, the benefits of averaging do not extend to the iterate X_n itself. If $a_n \rightarrow 0$ slowly enough, then over practical operating ranges there is not much difference between using a fixed ϵ or $a_n \rightarrow 0$. We work with a constant window of averaging and the constant gain case. The scheme will be seen to be robust. In Section IV it is shown that the use of averaging and feedback and a finite window does not hurt the iterate and yields an average that is $O(\epsilon^3)$ in typical cases. This is a vast improvement over the widely used constant gain case. In Section V, it is shown that the result is actually much better, and the result is quantified in terms of the feedback gain and the window of averaging. The considerable improvement that it can provide over the standard fixed gain case for both the direct iterate and the average are clearly seen from the estimates in Sections IV and V, and from the simulation results.

IV. ASYMPTOTIC ORDERS

Stability results and estimates of the orders of the asymptotic errors generally form the basis of asymptotic analysis of stochastic approximations. Such estimates are the subject of this section. Informal calculations suggest that the estimates (4.1b) and (4.1c) might be conservative. The stability estimates will be exploited in the next section, where we will get more precise asymptotics and a more explicit estimate of the errors as a function of the feedback gains and the window of averaging. As written, Theorem 4.1 shows that the use of feedback does not hurt the direct iterate and that the averaged iterate is much better than what one would get without feedback and averaging. In Sections V and VI, it will be seen that the direct iterate can be much better as well.

Assumptions: Define E_n as the conditional expectation given $\{X_0, \phi_i, \psi_i, i < n\}$. For real K_0 , we require, uniformly in n

$$\sum_{m=n}^{\infty} E_n \Phi_m \leq K_0, \quad (A4.1)$$

$$\sum_{m=n}^{\infty} E_n \rho_m \leq K_0. \quad (A4.2)$$

$$\{\Phi_i, \rho_i, i < \infty\} \text{ is bounded.} \quad (A4.3)$$

Note: Conditions such as (A4.1)–(A4.3) are quite convenient and not too restrictive in practice. They imply a certain rate of decrease of the correlation; i.e., a rate of decrease of the expectation of future “noise values” given the remote past. They are implied by various types of mixing conditions. See [3], [6] for examples. They do exclude the Gaussian case, but would include any reasonable truncation of a Gaussian sequence.

Theorem 4.1: Assume (A4.1)–(A4.3). Then, for (3.4a), (3.6)

$$\limsup E|X_n|^2 = O(\epsilon). \quad (4.1a)$$

$$\limsup_n E|\bar{X}_n|^2 = O(\varepsilon^2). \quad (4.1b)$$

If $E_n \rho_i = 0, i \geq n$, then

$$\limsup_n E|\bar{X}_n|^2 = O(\varepsilon^3). \quad (4.1c)$$

If (3.7) is used, then the right-hand sides of (4.1b), (4.1c) are replaced by $O(\varepsilon)$.

Comment: We have $E_n \rho_i = 0, i \geq n$, if the observation noises $\{\psi_n\}$ are mutually independent, have mean zero, and are independent of the $\{\phi_n\}$.

Proof: The proof will be divided into several parts. The basic technique is the perturbed Lyapunov function method [6], which has been found to be very useful for stability problems with non-Markovian systems. In Part 1, we try to apply a standard Lyapunov function method to the X_n , \bar{X}_n sequences and show that there is a problem due to certain "bad" terms. We then proceed to define the perturbed Lyapunov function, which is designed to deal with these bad terms. The procedure is straightforward, but a little tedious. The technique requires that a certain "fixed parameter" system be defined, and this is done in Part 2. Certain auxiliary functions (the perturbations) are defined in Part 3. In Part 4, the perturbed Lyapunov function is defined and appropriate estimates of the perturbation given.

Part 1: First, we use $b_n \rightarrow 0$ and attempt to use standard stochastic Lyapunov functions. It will be seen that these will not yield inequalities which can be used to prove the estimates (4.1). The problem concerns certain "bad" terms which appear and the inability of a classical Lyapunov function approach to exploit time scale differences. We will be able to cancel the effects of these bad terms via the use of suitable perturbations to the Lyapunov functions, which allow us to exploit time scale differences.

Define $C = [AI + G]^{-1}A$, and $Q = AI + G$. It is convenient to work first with a "centered" process defined by $\tilde{X}_n = X_n - C\bar{X}_n$. We can write

$$\tilde{X}_{n+1} = (I - cQ)X_n + c\xi_n + b_n C'(\bar{X}_n - X_n) \quad (4.2)$$

where the noise term ξ_n is defined by

$$\xi_n = \Phi_n X_n + \rho_n = \Phi_n[\tilde{X}_n + C'\bar{X}_n] + \rho_n. \quad (4.3)$$

We can write

$$\begin{aligned} E_n|\tilde{X}_{n+1}|^2 - |\tilde{X}_n|^2 &= 2\tilde{X}_n' E_n[\tilde{X}_{n+1} - \tilde{X}_n] + E_n|\tilde{X}_{n+1} - \tilde{X}_n|^2, \quad (4.4a) \\ 2\tilde{X}_n' E_n[\tilde{X}_{n+1} - \tilde{X}_n] &= -2\varepsilon\tilde{X}_n' Q\tilde{X}_n + 2\varepsilon\tilde{X}_n' E_n\xi_n + 2b_n\tilde{X}_n' C' \\ &\quad \times [\bar{X}_n - X_n]. \quad (4.4b) \end{aligned}$$

We now put (4.4b) into a form which will allow a better estimate to be obtained. For small positive k_1 , we will use the inequality

$$b_n|\tilde{X}_n||\bar{X}_n| \leq b_n \left[\frac{\varepsilon k_1}{b_n} |\tilde{X}_n|^2 + \frac{b_n}{\varepsilon k_1} |\bar{X}_n|^2 \right]. \quad (4.5)$$

The k_1 above and the positive real numbers k_2, k, q which will be introduced below can be as small as needed. But they

do not depend on ε, n . The purpose of inequalities such as (4.5) is to facilitate use of the time scale differences between \bar{X}_n and X_n , as will be seen. We use (4.5) in several of the following inequalities. If the term linear in k_1 is dominated by other terms for small k_1 , then we might omit it and adjust the orders appropriately.

Using the inequality (4.5) and writing $X_n = \tilde{X}_n + C\bar{X}_n$, yields

$$\begin{aligned} 2\tilde{X}_n' E_n[\tilde{X}_{n+1} - \tilde{X}_n] &\leq -2\varepsilon\tilde{X}_n' Q\tilde{X}_n + 2\varepsilon\tilde{X}_n' E_n\xi_n \\ &\quad + k_1 O(\varepsilon)|\tilde{X}_n|^2 + O\left(\frac{b_n^2}{\varepsilon}\right) \frac{1}{k_1} |\bar{X}_n|^2. \end{aligned}$$

Since we are concerned with asymptotic analysis, where $n \rightarrow \infty$ and ε is small, we always suppose (without loss of generality) that ε and the ratio b_n/ε are small, the latter going to zero with n . Then in all of the following estimates, we omit the terms which are dominated by others which appear; e.g., if for any "quantity" there is a term $O(\varepsilon) \times \text{quantity} + O(b_n) \times \text{quantity}$, we will omit $O(b_n) \times \text{quantity}$. From (4.4a), we can now write

$$\begin{aligned} E_n|\tilde{X}_{n+1}|^2 - |\tilde{X}_n|^2 &\leq -\varepsilon\tilde{X}_n' Q\tilde{X}_n + 2\varepsilon\tilde{X}_n' E_n\xi_n \\ &\quad + O\left(\frac{b_n^2}{\varepsilon}\right) \frac{1}{k_1} |\bar{X}_n|^2 + O(\varepsilon^2) E_n|\xi_n|^2. \quad (4.6) \end{aligned}$$

Using $X_n = \tilde{X}_n + C\bar{X}_n$, we can also write

$$\begin{aligned} |\bar{X}_{n+1}|^2 - |\bar{X}_n|^2 &= -2b_n|\bar{X}_n|^2 + 2b_n\bar{X}_n'(\tilde{X}_n + C'\bar{X}_n) \\ &\quad + O(b_n^2)(|\tilde{X}_n|^2 + |\bar{X}_n|^2). \quad (4.7) \end{aligned}$$

We will bound the right side of (4.7) by a more manageable function by using a form of (4.5), as follows. For small $k_2 > 0$, use

$$b_n|X_n||\bar{X}_n| \leq b_n \left[\frac{1}{k_2} |\tilde{X}_n|^2 + k_2 |\bar{X}_n|^2 \right]. \quad (4.8)$$

Again, when (4.8) is used, if the term linear in k_2 is dominated by other terms when k_2 is small, then we might omit it and adjust the orders appropriately. Now rewrite (4.7) as

$$\begin{aligned} |\bar{X}_{n+1}|^2 - |\bar{X}_n|^2 &\leq -2b_n\bar{X}_n'(I - C')\bar{X}_n \\ &\quad + O(b_n)k_2|\bar{X}_n|^2 + O(b_n)|X_n|^2 \frac{1}{k_2}. \quad (4.9) \end{aligned}$$

Part 2—Auxiliary Processes: Inequalities (4.6) and (4.9) cannot be used directly to get the moment estimates due to the appearance of certain "bad" terms. To deal with the "bad" term $2\varepsilon\tilde{X}_n' E_n\xi_n$ in (4.6), we need to exploit the time scale differences between the \tilde{X}_n and ξ_n sequences. This will be done by modifying the Lyapunov functions slightly. But first, we need to introduce various "fixed- r " processes. These are needed since the noise is not only correlated, but depends on the state. The approach is analogous to the one used successfully in [6]. First, we define some auxiliary sequences. In the next part of the proof, perturbations to the Lyapunov functions $|X_n|^2$ and $|\bar{X}_n|^2$ which use these auxiliary sequences will be defined and estimates of their values obtained. The proof will be completed in the Parts 4–6.

The noise ξ_n in (4.3) depends on the state X_n . Exploitation of the time scale differences is aided by the introduction of "fixed- x " noise process. This useful idea has wide applicability for getting limit results, whenever the noise is state dependent; e.g., see [6]. For each x , define the "fixed- x " noise process $\xi_n(x) = \Phi_n x + \rho_n$. For each n , define the noise sequence $\{\xi_m^n, m \geq n\}$ by

$$\xi_m^n = \Phi_m X_n + \rho_m. \quad (4.10)$$

Analogously, for each n and $m \geq n$ define the "fixed- x process" $\tilde{X}_m^n(x, x)$ by $\tilde{X}_m^n(r, x) = \tilde{x}$, and for $m \geq n$

$$\tilde{X}_{m+1}^n(x, \tilde{x}) = (I - \varepsilon Q) \tilde{X}_m^n(r, \tilde{x}) + \varepsilon \xi_m^n(x) \quad (4.11)$$

Note that x and \tilde{x} are parameters here. Define $\{X_m^n, m \geq n\}$ by $\tilde{X}_n^n = \tilde{X}_n$, and $X_m^n = X_m^n(X_n, \tilde{X}_n)$ for $m > n$ i.e., for $m \geq n$

$$\tilde{X}_{m+1}^n = (I - \varepsilon Q) \tilde{X}_m^n + \varepsilon \xi_m^n, \quad X_n^n = X_n \quad (4.12)$$

Thus, the parameter (x, \tilde{x}) in (4.11) is set to (X_n, X_n) , i.e., in $\{X_m^n, m \geq n\}$, the noise evolves (after time n) as though the state never changes, thus the term "fixed x " process.

Part 3 We will define various auxiliary sequences, which will be used for the Lyapunov function perturbations and obtain some estimates of their values. For each n , define the sequence

$$f_n^1 = \sum_{m=n}^{\infty} b_m E_n X_m^n \quad (4.13)$$

In the Appendix, it will be shown that

$$|f_n^1| = O\left(\frac{b_n}{\varepsilon}\right)(|X_n| + |\bar{X}_n|) + O(b_n) \quad (4.14a)$$

If $E_n \rho_n = 0$, $\varepsilon \geq n$, then we will get the estimate

$$|\hat{f}_n^1| = O\left(\frac{b_n}{\varepsilon}\right)(|X_n| + |\bar{X}_n|) \quad (4.14b)$$

To unify terminology, and write estimates such as (4.14a) and (4.14b) as one compact expression, we will use the symbol I : If we assume $E_n \rho_n = 0$, $\varepsilon \geq n$, then $I = 0$, otherwise it takes the value 1.

In the Appendix it is shown that the conditional difference satisfies

$$E_n f_{n+1}^1 - \hat{f}_n^1 = -b_n X_n + O\left(\frac{b_n^2}{\varepsilon} + b_n \varepsilon\right) \times (|\bar{X}_n| + |\tilde{X}_n|) + IO(b_n \varepsilon) \quad (4.15)$$

We have

$$E_n f_{n+1}^1 - \hat{f}_n^1 = -b_n X_n + \sum_{m=n+1}^{\infty} b_m E_n [\tilde{X}_m^{n+1} - \tilde{X}_m^n] \quad (4.16)$$

We now define one last perturbation. Define

$$\hat{f}_n^2 = 2\varepsilon \tilde{X}_n' \sum_{m=n+1}^{\infty} E_n \xi_m^n. \quad (4.17a)$$

Then

$$\hat{f}_n^2 = O(\varepsilon)(|\tilde{X}_n|^2 + |\bar{X}_n|^2) + IO(\varepsilon) \quad (4.17b)$$

where the estimate follows from (A4.1)–(A4.3). This perturbation will be used to cancel the $2\varepsilon \tilde{X}_n' E_n \xi_n$ term in (4.6). Now, evaluate the conditional difference (recall that $\xi_n^n = \xi_n$)

$$E_n \hat{f}_{n+1}^2 - \hat{f}_n^2 = -2\varepsilon \tilde{X}_n' E_n \xi_n + T_n^1 + T_n^2 \quad (4.18)$$

where we define

$$T_n^1 = 2\varepsilon E_n (\tilde{X}_{n+1} - \tilde{X}_n)' \sum_{m=n+1}^{\infty} E_{n+1} \xi_m^{n+1}$$

$$T_n^2 = 2\varepsilon \tilde{X}_n' \sum_{m=n+1}^{\infty} E_n [\xi_m^{n+1} - \xi_m^n].$$

We have

$$|T_n^1| + |T_n^2| = O(\varepsilon^2 + b_n \varepsilon) + O(\varepsilon^2 + b_n \varepsilon) \times (|X_n|^2 + |\bar{X}_n|^2) \quad (4.19)$$

Part 4 We are now prepared to define and use the perturbed stochastic Lyapunov functions. Define the perturbed Lyapunov function $V_n = |\tilde{X}_n|^2 + \hat{f}_n^2$. Then, via (4.6), (4.18), and cancelling terms where possible or where they are dominated by others

$$E_n V_{n+1} - V_n \leq -\frac{\varepsilon}{2} X_n' Q X_n + O\left(\frac{b_n^2}{\varepsilon} + \varepsilon^2\right) |\bar{X}_n|^2 \frac{1}{K_1} + O(\varepsilon^2) \quad (4.20)$$

Note, in particular, that the $-2\varepsilon \tilde{X}_n' E_n \xi_n$ in (4.18) cancels a term in (4.6). This is the essential reason for the form selected for \hat{f}_n^2 . Now, using the bound on the right side of (4.17), we see that there is some $K_1 > 0$ such that

$$E_n V_{n+1} - V_n \leq -\varepsilon K_1 V_n + O\left(\frac{b_n^2}{\varepsilon} + \varepsilon^2\right) |\bar{X}_n|^2 + O(\varepsilon^2). \quad (4.21)$$

Next, for $K_2 > 0$ define the Lyapunov function $V_n = K_2 b_n V_n / \varepsilon + |\tilde{X}_n|^2$. For large enough K_2 , there is $K_3 > 0$ such that (4.21), (4.9), and the fact that $I - C'$ is positive definite, imply that

$$E_n V_{n+1} - V_n \leq -b_n K_3 V_n + O(\varepsilon b_n) \quad (4.22)$$

Thus, taking expectations, iterating, and using the condition $\sum b_n = \infty$, yields

$$\limsup_n EV_n = O(\varepsilon) \quad (4.23)$$

Since $(K_2 b_n / \varepsilon) |f_n^2| = O(b_n)(|X_n|^2 + |\bar{X}_n|^2 + 1)$, for small ε and b_n we have

$$|\bar{X}_n|^2 \leq 2V_n + O(b_n)$$

Thus (4.23) implies that

$$\limsup_n E|\bar{X}_n|^2 = O(\varepsilon). \quad (4.24)$$

Now, taking expectations in (4.21), and substituting (4.24) into (4.21) yields

$$\limsup E(|X_n|^2 + \hat{f}_n^2) = O(\varepsilon)$$

Now, (4.1a) follows from this and the bound in (4.17).

Part 5: To date, we have mainly exploited the time scale differences between the $\{\tilde{X}_n, \bar{X}_n\}$ and $\{\Phi_n, \rho_n\}$ sequences. To improve the estimate (4.1a), we need to do a finer analysis. This will require the use of an additional perturbation, to exploit the time scale differences between the \tilde{X}_n and \bar{X}_n sequences. We get the estimates in this part and do the final Lyapunov function analysis in Part 6.

Define $\bar{f}_n^1 = 2\bar{X}_n' \tilde{f}_n^1$. By (4.14) and the inequality $|\bar{f}|O(b) \leq O(b) + O(b)|\bar{f}|^2$, we get

$$|\bar{f}_n^1| = O\left(\frac{b_n}{\varepsilon}\right)[|\tilde{X}_n|^2 + |\bar{X}_n|^2] + \hat{I}O(b_n). \quad (4.25)$$

We next evaluate the conditional difference, $E_n \bar{f}_{n+1}^1 - \bar{f}_n^1$. We can write

$$\begin{aligned} E_n \bar{f}_{n+1}^1 - \bar{f}_n^1 &= -2b_n \bar{X}_n' \tilde{X}_n \\ &\quad + 2E_n(\bar{X}_{n+1} - \bar{X}_n)' E_{n+1} \tilde{f}_{n+1}^1 \\ &\quad + 2\bar{X}_n' \sum_{m=n+1}^{\infty} b_m E_n[\tilde{X}_m^{n+1} - \tilde{X}_m^n] \\ &\equiv -2b_n \bar{X}_n' \tilde{X}_n + \bar{T}_n^1 + \bar{T}_n^2 \end{aligned} \quad (4.26)$$

where the \bar{T}_n^i are defined in the obvious way. From (4.14b),

$$\begin{aligned} |\bar{T}_n^1| &= O(b_n)(|\bar{X}_n| + |\tilde{X}_n|)O\left(\frac{b_n}{\varepsilon}\right)[|\bar{X}_n| + |\tilde{X}_n| + O(\varepsilon)] \\ &= O\left(\frac{b_n^2}{\varepsilon}\right)[|\tilde{X}_n|^2 + |\bar{X}_n|^2] + O(b_n^2). \end{aligned} \quad (4.27)$$

For $k > 0$ and $q > 0$ (which will be small numbers, but not going to zero with n or ε), we can write

$$\begin{aligned} |\bar{X}_n||\tilde{X}_n| &\leq \frac{k}{\varepsilon}|\bar{X}_n|^2 + \frac{\varepsilon}{k}|\tilde{X}_n|^2, \\ O(b_n\varepsilon)\hat{I}|\bar{X}_n| &\leq O(b_n\varepsilon)\hat{I}\left[\frac{q}{\varepsilon}|\bar{X}_n|^2 + \frac{\varepsilon}{q}\right]. \end{aligned}$$

Using these bounds and the estimates of the sum in (4.16) contained in the two right-hand terms of (4.15), we get

$$\begin{aligned} |\bar{T}_n^2| &= O\left(\frac{b_n^2}{\varepsilon} + b_n\varepsilon\right)\left(1 + \frac{k}{\varepsilon}\right)|\bar{X}_n|^2 \\ &\quad + O\left(\frac{b_n^2}{\varepsilon} + b_n\varepsilon\right)\frac{\varepsilon}{k}|\tilde{X}_n|^2 + \hat{I}O(b_n\varepsilon)\left[\frac{q}{\varepsilon}|\bar{X}_n|^2 + \frac{\varepsilon}{q}\right]. \end{aligned} \quad (4.28)$$

Part 6: We now improve the estimate (4.1a) and get (4.1b, c). Define the perturbed Lyapunov function $\bar{V}_n = |\bar{X}_n|^2 + \bar{f}_n^1$. Then, from (4.7), (4.26)–(4.28), and cancelling or dominating terms where possible, we get (where k, k_1 and q are as small as desired, but do not go to zero with n and ε)

$$\begin{aligned} E_n \bar{V}_{n+1} - \bar{V}_n &\leq -2b_n \bar{X}_n' (I - C) \bar{X}_n \\ &\quad + O\left(\frac{b_n^2}{\varepsilon}\right)(|\tilde{X}_n|^2 + |\bar{X}_n|^2) \\ &\quad + O(b_n^2) + O\left(\frac{b_n^2}{\varepsilon} + b_n\varepsilon\right)\left(\frac{k}{\varepsilon} + 1\right)|\bar{X}_n|^2 \\ &\quad + O\left(\frac{b_n^2}{\varepsilon} + b_n\varepsilon\right)\frac{\varepsilon}{k}|\tilde{X}_n|^2 + \hat{I}O(b_n\varepsilon)\left[\frac{q}{\varepsilon}|\bar{X}_n|^2 + \frac{\varepsilon}{q}\right]. \end{aligned}$$

Note, in particular, that the term $-2b_n \bar{X}_n' \tilde{X}_n$ in (4.26) cancels a similar term in (4.7). Taking expectations, dominating terms where possible, and using $\limsup E|\tilde{X}_n|^2 = O(\varepsilon)$ yields for small $\delta > 0$ (which goes to zero as $n \rightarrow \infty$ and $\varepsilon \rightarrow 0$)

$$\begin{aligned} E\bar{V}_{n+1} - \bar{V}_n &\leq -b_n E\bar{X}_n' (I - C - \delta I) \bar{X}_n \\ &\quad + O(b_n^2 + b_n\varepsilon^3) + \hat{I}O(b_n\varepsilon^2). \end{aligned}$$

For large enough n and small enough ε , there is $\delta_1 > 0$ such that

$$E\bar{V}_{n+1} - \bar{V}_n \leq -\delta_1 b_n E\bar{V}_n + O(b_n^2 + b_n\varepsilon^3) + \hat{I}O(b_n\varepsilon^2).$$

Thus

$$\limsup_n E\bar{V}_n = O(\varepsilon^3) + \hat{I}O(\varepsilon^2).$$

Now, for small b_n and ε

$$|\bar{X}_n|^2 \leq 2\bar{V}_n + O\left(\frac{b_n}{\varepsilon}\right)|\tilde{X}_n|^2 + O(b_n).$$

Equations (4.1b)–(4.1c) follow from (4.1a) and the last two equations.

The last assertion of the theorem (where $b_n = \alpha\varepsilon$) also follows from the above estimates.

V. RATES OF CONVERGENCE LIMITS OF NORMALIZED ITERATES

A. Stochastic Differential Equations for the Asymptotic Normalized Errors

In Theorem 4.1, it was shown that the use of feedback yields an asymptotic error of $O(\varepsilon)$ for X_n , and of $O(\varepsilon^2)$ or $O(\varepsilon^3)$ for \bar{X}_n . This does not tell the full story. Using these results, we will get a finer estimate of the asymptotic variances in this subsection. In the following subsection, similar results will be obtained for the time varying parameter case. We obtain weak convergence or rate of convergence results analogous to that assumed in (A2.1). The improvement in behavior due to feedback will become even clearer. Define $U_n = X_n/\sqrt{\varepsilon}$, $\bar{U}_n = \bar{X}_n/\sqrt{\varepsilon}$. Define the continuous parameter interpolations $U^\varepsilon(\cdot)$ by: $U^\varepsilon(t) = U_n$ for $t \in [n\varepsilon, (n+1)\varepsilon)$, and similarly define $\bar{U}^\varepsilon(\cdot)$. By (3.4a)

$$U_{n+1} = (I - \varepsilon Q)U_n + \sqrt{\varepsilon}\xi_n + \varepsilon A\bar{U}_n. \quad (5.1)$$

The methods in either [6] or in [8], [1] can be used to get the asymptotic distributions of both $U^\varepsilon(t)$ and $\bar{U}^\varepsilon(t)$ for large t . To simplify our work, we adapt the method of [6, Chapter 8, Section 3]. This requires the addition of two conditions which are quite unrestrictive. Recall that if $X_n = 0$ (the limit value), then $\xi_n = \phi_n \psi_n \equiv \rho_n$. Let us assume the following.

(A5.1): The sequence of processes $\sqrt{\varepsilon} \sum_{i=1}^{n+t/\varepsilon} \rho_i$, ($\rho_n = \phi_n \psi_n$) converges weakly to a Wiener process with covariance $R_0 = \sum_{i=1}^{\infty} E\rho_i \rho_i'$, as $\varepsilon \rightarrow 0$.

(A5.2): There are $n_\varepsilon \rightarrow \infty$ as $\varepsilon \rightarrow 0$ such that $\varepsilon n_\varepsilon \rightarrow$ and such that (in the sense of probability)

$$\frac{1}{n_\varepsilon} \sum_{i=n_\varepsilon}^{n+n_\varepsilon} \phi_i \phi_i' \rightarrow G,$$

$$\frac{1}{n_\varepsilon} \sum_{i=n_\varepsilon}^{n+n_\varepsilon} \phi_i \psi_i \rightarrow 0.$$

Comment on the Assumption: The second expression in (A5.2) is implied by (A5.1), but we keep it to help matching of the conditions with those in the reference. In the development in [6], the limit of the second sum in (A5.2) can be any constant vector. This simply creates a bias and can be used here too. (A5.2) is just a law of large numbers. Condition (A5.1) is also not very restrictive. It seems preferable to word the condition as it is since there is a very large literature on convergence of such sums to a Wiener process [6], [2], [3], and the convergence occurs under many different sets of more explicit and practical assumptions.

Theorem 5.1: Assume (A4.1)–(A4.3), (A5.1), (A5.2), and let $b_n \rightarrow 0$. Then there are $N_\varepsilon \rightarrow \infty$ such that $U^\varepsilon(\varepsilon N_\varepsilon + \cdot)$ converges weakly to the stationary solution of

$$-QUdt + R_0^{1/2}dW \quad (5.2)$$

where $W(\cdot)$ is a standard Wiener process.

If $b_n = \alpha\varepsilon$ for small ε , then Theorem 4.1 says that $\limsup_n E|\bar{X}_n|^2 = O(\varepsilon)$. But the actual situation is much better, and an analysis of this case gives additional insight into the case where $b_n \rightarrow 0$. The constant b_n case is of practical importance, since it is the case where there is slow "exponential forgetting," and is similar to the case where we average with a window of size $O(1/\varepsilon\alpha)$.

Theorem 5.2: Assume (A4.1)–(A4.3), (A5.1), (A5.2), and let $b_n = \alpha\varepsilon$ for small α . Then there are $N_\varepsilon \rightarrow \infty$ such that $(U^\varepsilon(\varepsilon N_\varepsilon + \cdot), \bar{U}^\varepsilon(\varepsilon N_\varepsilon + \cdot))$ converges weakly to the stationary solution of

$$dU = -QUdt + A\bar{U}dt + R_0^{1/2}dW$$

$$d\bar{U} = \alpha Udt - \alpha\bar{U}dt \quad (5.3)$$

where $W(\cdot)$ is a standard vector valued Wiener process.

Proof of Theorem 5.1: In view of the stability estimates of Theorem 4.1, the proof is essentially that in [6, Chapter 8, Section 3], and we only discuss the minor alterations. Note that our notation is different from that in the reference. Let $b_n \rightarrow 0$. The proof in [6] concerns the asymptotics without the feedback. But the only difference between that case and ours is the $\varepsilon A\bar{U}_n$ term in (5.1). Since $\limsup_n E|\bar{U}_n|^2 = O(\varepsilon)$ by Theorem 4.1, the added term does not contribute to the limit. The cited proof shows that there is a sequence of real numbers $t_\varepsilon \equiv \varepsilon N_\varepsilon \rightarrow \infty$ such that $U^\varepsilon(t_\varepsilon + \cdot) \Rightarrow U(\cdot)$ where $U(\cdot)$ is the stationary solution to (5.1), and R_0 is defined as in (A5.1).

Q.E.D.

Comment on Theorem 5.1: Loosely speaking, (5.2) implies that $X_n \sim N(0, \varepsilon V_0(Q))$, where $V_0(Q)$ is the stationary covariance of the solution to (5.2) and is

$$V_0(Q) = \int_0^\infty -Q^s R_0 e^{-Q^s} ds.$$

This means that the stationary covariance of U^ε goes down as A increases, being $O(1/A)$ for large A , i.e., for small fixed ε , as A increases the asymptotic covariance of X_n decreases. This is an immediate benefit of the feedback. Of course, one cannot let A grow without bound if ε is fixed since the algorithm will become unstable. We find in practice that the asymptotic results improve until near the point where stability does become a problem.

Proof of Theorem 5.2: Again, the proof is a mild extension of that in [6 Chapter 8, Section 3], and only a few comments will be made. We can write

$$U_{n+1} = (I - \varepsilon Q)U_n + \sqrt{\varepsilon}\xi_n + \varepsilon A\bar{U}_n, \quad (5.4)$$

$$\bar{U}_{n+1} = (1 - \alpha\varepsilon)\bar{U}_n + \varepsilon\alpha U_n. \quad (5.5)$$

Using the stability results in Theorem 4.1, the analysis in [6] can be readily extended to get that there is a sequence of real numbers $t_\varepsilon \rightarrow \infty$ such that

$$(U^\varepsilon(t_\varepsilon + \cdot), \bar{U}^\varepsilon(t_\varepsilon + \cdot)) \Rightarrow (U(\cdot), \bar{U}(\cdot))$$

where the limit process is the stationary solution to (5.3).

The Order of the Errors in Theorem 5.2 for Small α and Large: $\alpha \rightarrow A$. We will show that (for the stationary solution)

$$E|U(t)|^2 = O(1/A), \quad (5.6a)$$

$$E|\bar{U}(t)|^2 = O(\alpha/A). \quad (5.6b)$$

Thus, the asymptotic variances of $\bar{U}(t)$ is inversely proportional to the effective window of averaging, as for the case of Section II without feedback.

Proof of Claim: Since G is positive definite and symmetric, we can diagonalize (5.3). Write $G = P^{-1}DP$, $P^{-1} = P'$, where P is orthonormal and D is diagonal with positive entries. Let d_m denote the smallest diagonal entry in D . Define $V = PU$, $\bar{V} = P\bar{U}$, with coordinate variables V_i , etc. Then

$$dV = -(D + AI)Vdt + A\bar{V}dt + PR_0^{1/2}dW$$

$$d\bar{V} = \alpha(V - \bar{V})dt. \quad (5.8)$$

By Itô's Lemma and using $|R_0| = \text{trace } R_0$

$$\frac{dE|\bar{V}_i|^2}{dt} = 2\alpha(EV_i\bar{V}_i - E\bar{V}_i^2), \quad (5.9a)$$

$$\frac{dE|V|^2}{dt} = -2AE|V|^2 - 2E\bar{V}'DV + 2AEV'\bar{V} + |R_0|, \quad (5.9b)$$

$$\frac{dEV'\bar{V}}{dt} = \alpha E|V|^2 - \alpha V'\bar{V} - E\bar{V}'DV$$

$$+ A(E|\bar{V}|^2 - E\bar{V}'V). \quad (5.9c)$$

The deterministic part of the system (5.8) is stable. For the stationary solution, we have

$$EV_i\bar{V}_i = E\bar{V}_i^2, \quad (5.10a)$$

$$2AE|V|^2 + 2EV'\bar{V} = 2AE|\bar{V}|^2 + |R_0|. \quad (5.10b)$$

$$\alpha E|\bar{V}|^2 + E\bar{V}'DV = \alpha E|V|^2. \quad (5.10c)$$

In (5.10c), we used (5.10a) to cancel the coefficient of A in the right side of (5.9c). Using (5.10a) again and taking lower

bounds in (5.10b) and (5.10c) yields

$$(d_m + \alpha)E|\bar{V}|^2 \leq \alpha E|V|^2, \quad (5.11)$$

$$2(A + d_m)E|V|^2 \leq 2AE|\bar{V}|^2 + |R_0| \quad (5.12)$$

from which we get (5.6).

B. Tracking Time Varying Parameters

Feedback and averaging can also be used when the parameters to be tracked vary with time. Extensive analyses of tracking algorithms are in [1], [4]. We view the main practical applications of averaging with feedback to be for constant parameter systems, and the applications to communications, parameter estimation, adaptive noise cancellation and elsewhere in signal processing cited earlier. But, for robustness considerations it is important to know that the basic behavior is continuous under small time variations (small β below). The dependence on β is seen in (5.16). If $\beta = o(\alpha)$, then (5.16) can be improved.

Let $b_n = \alpha\epsilon$, and let θ_n denote the value of the parameter (replacing θ) at time n . For β a real number, suppose that the parameter evolves according to

$$\theta_{n+1} = \theta_n + \beta\epsilon w_n.$$

Here, $\{w_n\}$ is a driving sequence of bounded random variables with mean zero, independent of $\{\phi_n, \psi_n\}$ and satisfying

$$\sum_{m=0}^{\infty} E_n w_m \leq K_0 \quad (A5.3)$$

for some $K_0 < \infty$ and all n , a condition analogous to (A4.1). Note that we do not require $\{w_n\}$ to be mutually independent. We will also use the conditions [the analogs of (A5.1), (A5.2)]

(A5.4): The sequence of processes $\sqrt{\epsilon} \sum_{i=n}^{n+t/\epsilon} w_i$ converges weakly to a Wiener process with covariance matrix $R_1 = \sum_{i=0}^{\infty} E w_i w_i^T$.

(A5.5): There is a sequence of real numbers $n_\epsilon \rightarrow \infty$ as $\epsilon \rightarrow 0$ such that $\epsilon n_\epsilon \rightarrow 0$ and (in the sense of probability)

$$\frac{1}{n_\epsilon} \sum_{i=0}^{n+n_\epsilon} w_i \rightarrow 0.$$

The Evolution Equation: Redefine $X_n = \hat{\theta}_n - \theta_n$, $\bar{X}_n = \bar{\theta}_n - \theta_n$, where

$$\bar{\theta}_n = (I - \alpha\epsilon)\bar{\theta}_{n-1} + \alpha\epsilon\hat{\theta}_{n-1}.$$

Then (the analog of (3.4a))

$$X_{n+1} = (I - \epsilon G)X_n + \epsilon\xi_n + \epsilon A(\bar{X}_n - X_n) + (\theta_n - \theta_{n+1})$$

where ξ_n is defined as in Section III, and

$$\bar{X}_{n+1} = (1 - \alpha\epsilon)\bar{X}_n + \alpha\epsilon X_n + (\theta_n - \theta_{n+1}).$$

The entire analysis for the non time-varying case can be carried over to the present case with only trivial changes. By a proof almost identical to that of Theorem 4.1, we have the following result.

Theorem 5.3: Assume (A4.1)–(A4.3) and (A5.1)–(A5.3). Then for small β and α

$$\limsup_n E|X_n|^2 = O(\epsilon),$$

$$\limsup_n E|\bar{X}_n|^2 = O(\epsilon).$$

With Theorem 5.3 in hand, precise analogs of Theorem 5.2 can be obtained by exactly the same analysis, and we will only write the appropriate formulation. Redefine $U_n = X_n/\sqrt{\epsilon}$, $\bar{U}_n = \bar{X}_n/\sqrt{\epsilon}$. Then

$$U_{n+1} = (I - \epsilon Q)U_n + \sqrt{\epsilon}\xi_n + \epsilon A\bar{U}_n - \sqrt{\epsilon}\beta w_n, \quad (5.13)$$

$$\bar{U}_{n+1} = (1 - \alpha\epsilon)\bar{U}_n + \epsilon\alpha U_n - \sqrt{\epsilon}\beta w_n. \quad (5.14)$$

Then the analysis which led to Theorem 5.2 and the order estimates which followed yield the following.

Theorem 5.4: Assume (A4.1)–(A4.3) and (A5.1)–(A5.5). Then, for small β and α , there are $N_\epsilon \rightarrow \infty$ such that $(U^\epsilon(\epsilon N_\epsilon + \cdot), \bar{U}^\epsilon(\epsilon N_\epsilon + \cdot))$ converges weakly to the stationary solution of

$$dU = -QUdt + A\bar{U}(t)dt + R_0^{1/2}dW + \beta R_1^{1/2}dW_1$$

$$d\bar{U} = \alpha Udt - \alpha\bar{U}dt + \beta R_1^{1/2}dW_1 \quad (5.15)$$

where $W(\cdot)$ and $W_1(\cdot)$ are mutually independent vector valued Wiener processes. If $\beta = O(\alpha)$, then (for the stationary solution)

$$E|U(t)|^2 = O\left(\frac{A\beta^2}{\alpha}\right) + O(1/A), \quad (5.16a)$$

$$E|\bar{U}(t)|^2 = O\left(\frac{A\beta^2}{\alpha}\right) + O(\alpha/A). \quad (5.16b)$$

VI. SIMULATION DATA

Numerous simulations have supported the theoretical conclusions, and the following tables are typical of the runs taken. The simulations were made with a five-dimensional system, where the five-dimensional vector ϕ_n was computed as $\phi_n = (\phi_n, \phi_{n-1}, \phi_{n-2}, \phi_{n-3}, \phi_{n-4})$, where the sequence of random variables ϕ_n is obtained from the dynamical equation $\phi_{n+1} = \phi_n/2 + \zeta_n$ where ζ_n is a sequence of mutually independent Gaussian random variables with mean zero and variance 1.0. The parameter θ is fixed at $\theta = (4.0, -4.2, 3.0, 2.7, -3.0)$. The standard deviation of the observation noise ψ_n was 6.0. The noise level was chosen high enough to allow us to distinguish the general differences among the algorithms, since for small noise levels all the algorithms worked well. If one wishes to distinguish between the small errors at low observation noise levels, the nonfeedback algorithm is often preferable. Asymptotically, the best values of ϵ are very small and decrease as the run size increases. But for run lengths of fixed finite size, the performance will degenerate if ϵ becomes too small. Generally, we have tried to compare the performance of the algorithms with and without feedback using the parameters at which each works well. The actual averaged value that was used in the (both feedback and

TABLE I

 $\epsilon = .05, A = 5, \text{WINDOW} = 150$

case		parameter#				
		1	2	3	4	5
no FB	X_n	1.133	1.212	1.173	1.251	1.014
no FB	\bar{X}_n	.230	.297	.257	.370	.207
FB	X_n	.275	.238	.216	.278	.255
FB	\bar{X}_n	.065	.037	.019	.086	.060

TABLE II

 $\epsilon = .05, A = 22, \text{WINDOW} = 150$

case		parameter#				
		1	2	3	4	5
no FB	X_n	1.133	1.212	1.173	1.251	1.014
no FB	\bar{X}_n	.230	.297	.257	.370	.207
FB	X_n	.206	.191	.217	.299	.325
FB	\bar{X}_n	.045	.014	.020	.083	.095

TABLE III

 $\epsilon = .1, A = 2.5, \text{WINDOW} = 150$

case		parameter#				
		1	2	3	4	5
no FB	X_n	6.672	7.633	8.070	5.780	6.351
no FB	\bar{X}_n	1.126	.968	.949	.757	6.1
FB	X_n	2.266	2.820	3.334	2.974	4.409
FB	\bar{X}_n	.327	.146	.202	.214	.142

TABLE IV

 $\epsilon = .02, A = 9.5, \text{WINDOW} = 300$

case		parameter#				
		1	2	3	4	5
no FB	X_n	0.350	0.441	0.385	0.489	0.350
no FB	\bar{X}_n	.086	.125	.085	.194	.088
FB	X_n	.061	.105	.131	.067	.096
FB	\bar{X}_n	.013	.064	.090	.027	.059

TABLE V

 $\epsilon = .05, A = 5, \text{WINDOW} = 300$

case		parameter#				
		1	2	3	4	5
no FB	X_n	1.133	1.212	1.173	1.251	1.014
no FB	\bar{X}_n	.124	.131	.096	.225	.110
FB	X_n	.274	.229	.245	.239	.242
FB	\bar{X}_n	.049	.016	.037	.038	.037

TABLE VI

 $\epsilon = .05, A = 22, \text{WINDOW} = 300$

case		parameter#				
		1	2	3	4	5
no FB	X_n	1.133	1.212	1.173	1.251	1.014
no FB	\bar{X}_n	.124	.131	.096	.225	.110
FB	X_n	.246	.197	.329	.243	.320
FB	\bar{X}_n	.064	.007	.114	.010	.070

nonfeedback) algorithm was

$$\bar{X}_n = \frac{1}{\text{window}} \sum_{n-\text{window}}^{n} X_i$$

The averaging started at iteration 50, using a growing window of averaging until the desired window size was reached, after which we averaged only over the desired window size. The computation of the mean square errors commenced at the 300th iteration so as to minimize the effect of the transient period. The tables list the sample mean square errors at 5000 iterations; namely

$$\frac{1}{5000 - 300} \sum_{n=300}^{5000} |\theta'_n - \theta'|^2$$

where θ' is the i th component of θ . The sum is quite stable and the values were essentially the same after 1000 iterations. Hence, they can be taken for good approximations to the actual mean square errors. Generally, the algorithm with feedback is not as sensitive to the value of ϵ and can work well at somewhat larger values of ϵ , although ϵ cannot be made too large without causing stability problems. This ability to work well with somewhat larger ϵ is a considerable advantage, in that it yields a shorter transient period.

In all cases, both with and without feedback, it can be seen that averaging yields substantial improvements. This was true even when the original iterates X_n were poor, as in Table III (and also Tables I, II, V, VI, to some extent) for the algorithm without feedback. Clearly, as expected, the performance improves as the window of averaging increases. Refer to Table III, where ϵ and $A\epsilon$ are too large for good performance. But even there, with the poor values given by the X_n , averaging was a distinct advantage and more so for the feedback case.

If the observation noise level is not too small, then the averaged iterate \bar{X}_n for the feedback case is generally better than that for the nonfeedback case and is often much better as seen from the tables. For such a case, the errors X_n are also generally smaller in the feedback case. This is important if

one wishes a good "on-line" algorithm. Indeed, in some cases the errors X_n with feedback are as good as the errors \bar{X}_n without feedback. While the last word has not been said on the general family of algorithms considered and more testing and comparison needs to be done (as is usual for a new idea), it is clear that the use of feedback is quite promising and has many advantages, particularly for an algorithm where one wants a good estimate "on line."

We also note that the use of feedback makes the behavior of the X_n less sensitive to the value of G , as seen from the forms of the stochastic differential equations derived in Section V.

APPENDIX

DETAILS OF ESTIMATES IN THEOREM 4.1

Proof of (4.14) By definition, for $m \geq n$

$$\hat{X}_m^n = (I - \epsilon Q)^{m-n} \hat{X}_n + \sum_{i=n}^{m-1} \epsilon (I - \epsilon Q)^{m-i-1} \xi_i^n. \quad (\text{A.1a})$$

We will also use the representation [(A 1a), but starting at time $n + 1$]

$$\tilde{X}_m^n = (I - \varepsilon Q)^{m-n-1} \tilde{X}_{n+1}^n + \sum_{i=n+1}^m \varepsilon (I - \varepsilon Q)^{m-i-1} \xi_i^n \quad (\text{A } 1b)$$

By (A 1a)

$$\begin{aligned} \tilde{f}_n^1 &= \sum_{m=n}^{\infty} b_m (I - \varepsilon Q)^{m-n} X_n \\ &+ \sum_{m=n}^{\infty} \sum_{i=n}^{m-1} b_m \varepsilon (I - \varepsilon Q)^{m-i-1} E_n \xi_i^n \end{aligned} \quad (\text{A } 2)$$

Using the fact that $\sum_{m=0}^{\infty} (I - \varepsilon Q)^m = O(1/\varepsilon)$, the first term on the right side of (A 2) is $O(b_n/\varepsilon)|X_n|$. Using (A4.1)–(A4.3) and the definition (4.10) to evaluate the second term on the right side of (A 2) yields $O(b_n)(|\tilde{X}_n| + |\bar{X}_n|) + \tilde{I}O(b_n)$. Thus (4.14) holds.

Proof of (4.15) Using (A4.1)–(A4.3) and the representation for \tilde{X}_m^n from (A 1b) and for X_m^{n+1} from (A 1a) (with $n+1$ replacing n there), the component of the sum in (4.16) which is due to the “propagation of the initial conditions” (rather than due to the noise sequences $\{\xi_m^n, \xi_m^{n+1}, m \geq n+1\}$) can be seen to be

$$\begin{aligned} E_n \sum_{m=n+1}^{\infty} b_m (I - \varepsilon Q)^{m-n-1} X_{n+1} \\ - \sum_{m=n+1}^{\infty} b_m (I - \varepsilon Q)^{m-n-1} X_{n+1} \end{aligned}$$

Since

$$\begin{aligned} X_{n+1} &= (I - \varepsilon Q) \tilde{X}_n + \varepsilon \xi_n + b_n C'(\bar{X}_n - X_n) \\ &= \tilde{X}_{n+1}^n + b_n C'(\bar{X}_n - X_n) \end{aligned}$$

the above difference equals

$$\sum_{m=n+1}^{\infty} b_m (I - \varepsilon Q)^{m-n-1} b_n C'(\bar{X}_n - X_n) \quad (\text{A } 3)$$

which equals

$$O\left(\frac{b_n^2}{\varepsilon}\right)(|\tilde{X}_n| + |\bar{X}_n|) \quad (\text{A } 4)$$

The component of the sum in (A 4) due to the “noise” $\{\xi_m^n, \xi_m^{n+1}, m \geq n+1\}$ is

$$\sum_{m=n+1}^{\infty} \sum_{i=n+1}^{m-1} b_m \varepsilon (I - \varepsilon Q)^{m-i-1} E_n [\xi_i^{n+1} - \xi_i^n]$$

Noting that $\xi_i^{n+1} - \xi_i^n$ equals $\Phi_i(X_{n+1} - X_n)$, and using (A4.1)–(A4.3), we get the order [recall the definition of I below (4.14b)]

$$O(b_n \varepsilon + b_n^2)(|\tilde{X}_n| + |\bar{X}_n|) + IO(b_n \varepsilon) \quad (\text{A } 5)$$

(A 4) and (A 5) yield the right side of (4.15)

REFERENCES

- [1] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximation*. New York: Springer-Verlag, 1990.
- [2] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968.
- [3] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*. New York: Wiley, 1986.
- [4] L. Guo, L. Ljung, and P. Priouret, ‘Tracking performance analyses of the forgetting factor RLS algorithm’, in *Proc. 31st Conf. Decis. Contr.*, Tucson, AZ, 1992, pp. 688–693.
- [5] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1990.
- [6] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic System Theory*. Cambridge, MA: MIT Press, 1984.
- [7] H. J. Kushner and D. S. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*. New York: Springer-Verlag, 1978.
- [8] H. J. Kushner and A. Schwartz, ‘Weak convergence and asymptotic properties of adaptive filters with constant gains’, *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 177–182, 1984.
- [9] H. J. Kushner and J. Yang, ‘Stochastic approximation with averaging: Optimal asymptotic rates of convergence for general processes’, *SIAM J. Contr. Optim.*, vol. 31, pp. 1045–1062, 1993.
- [10] H. J. Kushner and G. Yin, ‘Asymptotic properties of distributed and communicating stochastic approximation algorithms’, *SIAM J. Contr. Optim.*, vol. 25, pp. 1266–1290, 1987.
- [11] B. T. Polyak and A. B. Juditsky, ‘Acceleration of stochastic approximation by averaging’, *SIAM J. Contr. Optim.*, vol. 30, pp. 838–855, 1992.
- [12] D. Ruppert, ‘Efficient estimators from a slowly convergent Robbins-Munro process’, School of Operations Research and Industrial Engineering, Cornell Univ., Tech. Rep. 787, 1988.
- [13] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.
- [14] G. Yin, ‘On extensions of Polyak’s averaging approach to stochastic approximation’, *Stochastics*, vol. 36, pp. 245–264, 1992.
- [15] G. Yin, *Stochastic Approximation via Averaging: Polyak’s Approach Revisited* (Lecture Notes in Economics and Mathematical Systems). Berlin: Springer-Verlag, vol. 374, 1992, pp. 119–134.



Harold Kushner (S 54–A 56–M 59 SM 73 F 74) received the B.S.E.E. degree from City College of New York, NY, in 1955 and the Ph.D. degree in electrical engineering from the University of Wisconsin-Madison in 1958.

He is Director of the Lefschetz Center for Dynamical Systems, Brown University, Providence, RI. His current research interests include large deviations, wide bandwidth systems, weak convergence methods, systems approximations, and limit theorems, singularly perturbed systems, and heavy traffic approximations.

Dr. Kushner has won the IEEE Field Award in Control Systems and the Franklin Institute Louis F. Levy medal. He has written seven books and over 150 papers on virtually all aspects of stochastic systems.



Jichuan Yang was born in Sichuan, China, in 1963. He received the B.S. degree in mathematics from Qinghua University, P.R. China, in 1985 and the Ph.D. degree in applied mathematics from Brown University, Providence, RI, in 1991.

Dr. Yang is currently working in a postdoctoral position in the Division of Applied Mathematics, Brown University. His current research interests include numerical methods in stochastic control, optimization, stochastic approximation algorithms, and financial derivative securities pricing models.

Exponential Stabilization of Nonholonomic Chained Systems

O. J. Sørдалen, *Member, IEEE*, and O. Egeland, *Member, IEEE*

Abstract—This paper presents a feedback control scheme for the stabilization of two-input, driftless, chained nonholonomic systems, also called chained form. These systems are controllable but not asymptotically stabilizable by a smooth static-state feedback control law. In addition, exponential stability cannot be obtained with a smooth, time-varying feedback control law. Here, global, asymptotical stability with exponential convergence is achieved about any desired configuration by using a nonsmooth, time-varying feedback control law. The control law depends, in addition to the state and time, on a function which is constant except at predefined instants of time where the function is recomputed as a nonsmooth function of the state. The inputs are differentiable with respect to time and tend exponentially toward zero. For use in the analysis, a lemma on the exponential convergence of a stable time-varying nonlinear system perturbed by an exponentially decaying signal is presented. Simulation results are also shown.

I. INTRODUCTION

NONHOLONOMIC chained systems can be used to represent a large class of mechanical systems. Important and well-known examples are unicycles, four-wheeled cars and n -trailer systems. Accordingly, control schemes for nonholonomic chained systems has a potential of being applied to a large number of mechanical systems. The problem of designing stabilizing feedback controllers for nonholonomic chained systems is a challenging one since the system is not stabilizable by a smooth static-state feedback law [4]. Moreover, the problem of exponential stabilization, which is the topic of this paper, cannot be solved by any smooth feedback law [10].

The nonholonomic chained system considered in this paper is the chained form [26]

$$\begin{aligned}\dot{x}_1 &= u_1 \\ \dot{x}_2 &= u_2 \\ \dot{x}_i &= x_{i-1}u_1, \quad i \in \{3, \dots, n\}.\end{aligned}$$

A constructive procedure to transform a nonholonomic system with two inputs into this chained form was presented by [25] under certain conditions on the input vectors. Necessary and

sufficient conditions for converting a nonholonomic system into chained form were derived in [23] based on the theory of exterior differential systems using the Goursat normal form theorem. One of the results from this work is that all two-input, regular nonholonomic systems in three and four dimensions are locally feedback equivalent to a nonholonomic chained form. In [33], the kinematic model of a car with n trailers was converted into chained form. This conversion was global in the position of the system and local in the orientations. Another conversion into chained form was proposed in the framework of exterior differential systems in [35] based on [33], [23], and [30]. This conversion of the kinematics of the n -trailer system allowed any orientation of the last trailer. Control strategies for nonholonomic chained systems can, therefore, be used for the control of a broad class of nonholonomic, mechanical systems.

The control of chained form and of more general nonholonomic systems is a very active field of research. The problem of nonholonomic motion planning was introduced by [17] who proved that a car-like robot with one nonholonomic constraint is controllable. Open-loop planners for low-dimensional mobile robots have been proposed in [18], [1], [19]. Because of the invertible transformation in [33], these planners can also be used to plan a path for low-dimensional chained systems. Other open-loop strategies have explored control theoretic approaches using differential geometry tools to control nonholonomic systems. Sinusoids were proposed by [24] to steer in open-loop nonholonomic systems on a special canonical form including chained form. A generalization of the use of sinusoidal inputs to generate motion at a given level of Lie brackets of the input vectors was given in [16], [15], [12], [22] for nilpotent and nilpotentizable systems using an extended system with additional input vectors corresponding to higher order Lie brackets of the original system. Since the nonholonomic chained system is nilpotent, these open-loop strategies can also be used to steer such a system in open loop.

To make the control more robust with respect to disturbances and errors in the initial condition, some stabilizing closed-loop approaches have also been proposed. Characteristic for the nonholonomic systems encountered in robotics is that they cannot be stabilized by a smooth static-state feedback control law which has been shown with Brockett's theorem [4]. This has been further discussed by [2] and [8] for nonholonomic mechanical systems. Therefore, a discontinuous feedback control law was proposed by [5] to make the kinematic model of a three-dimensional robot globally, exponentially converge to a given configuration. This system is equivalent to a three-dimensional chained form. Another

Manuscript received July 2, 1993; revised March 15, 1994. Recommended by Associate Editor, A. M. Bloch. This work was supported in part by the Royal Norwegian Council for Scientific and Industrial Research (NTNF) and the Center of Maritime Control Systems at NTH/SINTEF.

O. J. Sørдалen was with Department of Engineering Cybernetics, The Norwegian Institute of Technology, 7034 Trondheim-NTH, Norway and is now with ABB Corporate Research Center, ABB Teknolog AS, N-1361 Billigstad, Norway.

O. Egeland is with the Department of Engineering Cybernetics, The Norwegian Institute of Technology, 7034 Trondheim-NTH, Norway.

IEEE Log Number 9406097.

discontinuous feedback approach was proposed by [3] for the same system with acceleration inputs instead of velocity inputs. This approach makes the system reach the origin in finite time in the case of no disturbances.

The use of time-varying feedback control laws is another approach to stabilize nonholonomic systems about a constant configuration. This approach was first studied by [30] for the stabilization of a cart. This approach was further developed in [29] for a car-like mobile robot with a steering wheel. This system is locally equivalent to a four-dimensional chained form. Constructive approaches were presented in [31], [28]. The existence of stabilizing time-varying feedback control laws for more general nonholonomic systems was studied in [6]. The design methods in [28] were extended in [7] to the more general situation given in [6]. An algorithm for computing time-periodic feedback solutions for nonholonomic motion planning was presented in [10]. This approach considered the extended system using Lie bracket completion vectors such as in the open-loop strategy of [16]. This feedback algorithm was based on multi-scaling averaging techniques and highly oscillatory inputs. These time-dependent approaches also work for chained form. Another time-varying smooth feedback control law to stabilize chained form was derived in [34] for power forms based on the work on the use of sinusoids.

These smooth time-varying feedback control laws yield asymptotic stability but not exponential since time-periodic smooth feedback cannot be exponentially stabilizing [9], [10]. To improve the rate of convergence, a feedback control law was proposed by [27] to obtain local exponential convergence to a neighborhood of the origin for a three-dimensional chained form. The asymptotic behavior was obtained by letting the control law be time varying. The exponential convergence to the neighborhood was obtained by letting the control law be nonsmooth at the origin. No feedback control law, however, has been presented in previous work that exponentially stabilizes a nonholonomic chained form of an arbitrary dimension about any constant configuration.

In this paper, the problem of exponential stabilization is addressed and a new feedback approach is proposed for chained form. The stabilization is achieved by letting the state feedback control law depend on time and on a function which is nonsmooth with respect to the state at discrete instants of time. The proposed feedback control law globally stabilizes the system about the origin with exponential convergence. Asymptotic stability about any desired configuration is obtained by using a coordinate transformation. The resulting closed-loop system is not exponentially stable as defined by [14, p. 168], but it is shown to have a property which will be termed \mathcal{K} -exponential stability.

The paper is organized as follows: The concept of \mathcal{K} -exponential stability is proposed in Section II. A lemma on the exponential convergence of a nonlinear time-varying system which is perturbed by an exponentially decaying signal is given in Section III. The nonholonomic chained system is presented in Section IV. The control law is presented in Section V. The convergence of a part of the system is analyzed in Section VI. The stability of the total system is analyzed in Section VII. In Section VIII, a coordinate transformation

is presented such that a control law to control the chained form to the zero configuration can be used for the control to any desired configuration. The stabilizing feedback strategy is illustrated by a simulation example in Section IX. The conclusions are given in Section X.

II. \mathcal{K} -EXPONENTIAL STABILITY

In this section, the concept of \mathcal{K} -exponential stability is introduced. It will be shown in later sections that the controller proposed in this paper makes the chained system \mathcal{K} -exponentially stable. First, we need the following notion [11, Definition 2.5].

Definition 1: A continuous function $\alpha: R^+ \rightarrow R^+$ is said to be of class \mathcal{K} (or belong to class \mathcal{K}) if it is strictly increasing and $\alpha(0) = 0$.

One property of class \mathcal{K} functions is that, [14] and [11]

$$\alpha_1, \alpha_2 \in \text{class } \mathcal{K} \Rightarrow \alpha_1 \circ \alpha_2 \in \text{class } \mathcal{K}. \quad (1)$$

We then define the following.

Definition 2: Consider the nonlinear, time-varying system

$$\dot{x} = f(x, t) \quad x \in D \subset R^n, \quad t \geq t_0. \quad (2)$$

System (2) is \mathcal{K} -exponentially stable about x_p iff there exist a neighborhood $\Omega_p \subset D$ about x_p , a positive constant λ , and a function $h(\cdot)$ of class \mathcal{K} such that all solutions $x(t)$ of (2) satisfy

$$\forall x(t_0) \in \Omega_p \forall t \geq t_0, \|x(t) - x_p\| \leq h(\|x(t_0) - x_p\|)e^{-\lambda(t-t_0)} \quad (3)$$

where the constant λ and the neighborhood Ω_p are independent of t_0 , and $\|\cdot\|$ denotes a norm in R^n .

If (3) is satisfied for $\Omega_p = D$, then system (2) is globally, \mathcal{K} -exponentially stable about x_p . According to this definition, if system (2) is \mathcal{K} -exponentially stable at x_p , then it is uniformly asymptotically stable as defined by e.g., [14, 4.3], and in addition it has an exponential rate of convergence.

The concept of \mathcal{K} -exponential stability is called exponential stability in [21]. The term exponential stability is however often used for the special case where the function $h(\cdot)$ is linear, [11] and [14], i.e.,

$$h(\|x(t_0) - x_p\|) = r\|x(t_0) - x_p\|.$$

Here, r is a positive constant independent of t_0 and $x(t_0)$. This means that \mathcal{K} -exponential stability corresponds to a weaker form of stability than the usual concept of exponential stability. \mathcal{K} -exponential stability and exponential stability are equal, however, with respect to the rate of convergence. Therefore, the notion "exponential stabilization" is used in the title of this paper; although, only \mathcal{K} -exponential stability will be proved.

Locally, a similar definition of exponential stability was introduced in [13] in terms of homogeneous norms. This concept has been used in [20] to investigate the convergence rates for controllers for low dimensional nonholonomic systems in so-called power form which is equivalent to chained form, [34].

III. A LEMMA ON EXPONENTIAL CONVERGENCE

The following lemma is useful for establishing exponential convergence for a class of time-varying systems. It will be used in the convergence analysis of the control law.

Lemma 1: Consider the nonlinear, one-dimensional, time-varying system

$$\dot{x} = -a(x, t)x + d(x, t), \quad t \geq t_0, \quad x(t_0) \in R \quad (4)$$

under the following assumptions:

- There exists a solution $x(t)$ of (4) for any $x(t_0)$ and any $t \geq t_0$; see Remark 2 below.
- $a(x, t)$ has the property that for all $x(t)$

$$\left| \int_{t_0}^t (a(x(\tau), \tau) - \lambda) d\tau \right| \leq P, \quad \forall t \geq t_0 \quad (5)$$

where λ and P are positive constants.

- The signal $d(x, t)$ is bounded for any $t \geq t_0$ and any $x(t)$ by

$$|d(x(t), t)| \leq D e^{-\gamma(t-t_0)} \quad (6)$$

for some positive constants D and γ .

Then

$$\forall \epsilon > 0, \quad |x(t)| \leq \epsilon (|x(t_0)| + D) e^{-(\alpha - \epsilon)(t-t_0)}$$

where

$$\alpha = \min\{\lambda, \gamma\} > 0, \quad \epsilon = \max\left\{\epsilon^P, \frac{\epsilon^{2P}}{\epsilon}\right\}.$$

Proof We denote

$$F(t) = \int_{t_0}^t a(x(\tau), \tau) d\tau$$

where $x(t)$ is a solution of (4). Multiplying (4) with $e^{F(t)}$ gives

$$\frac{d}{dt}(x(t)e^{F(t)}) = d(x(t), t)e^{F(t)}.$$

This implies

$$x(t)e^{F(t)} = x(t_0) + \int_{t_0}^t d(x(\tau), \tau)e^{F(\tau)} d\tau.$$

Dividing by $e^{F(t)}$ then gives the following (implicit) expression for $x(t)$

$$x(t) = e^{-F(t)} \left(x(t_0) + \int_{t_0}^t e^{F(\tau)} d(x(\tau), \tau) d\tau \right). \quad (7)$$

Property (5) implies that

$$|F(t) - \lambda(t - t_0)| \leq P \quad (8)$$

which is equivalent to

$$-P + \lambda(t - t_0) \leq F(t) \leq P + \lambda(t - t_0). \quad (9)$$

By using (7), (9), and (6) we get

$$|x(t)| \leq e^P e^{-\lambda(t-t_0)} \left(|x(t_0)| + D \int_{t_0}^t e^P e^{(\lambda-\gamma)(\tau-t_0)} d\tau \right).$$

In the case that $\lambda = \gamma$ we find

$$|x(t)| \leq e^P |x(t_0)| e^{-\lambda(t-t_0)} + D e^{2P} (t - t_0) e^{-\lambda(t-t_0)}$$

In the case that $\lambda \neq \gamma$ we find

$$|x(t)| \leq e^P |x(t_0)| e^{-\lambda(t-t_0)} + \frac{D e^{2P}}{\lambda - \gamma} (e^{-\gamma(t-t_0)} - e^{-\lambda(t-t_0)})$$

Since

$$1 - e^{-b} \leq b, \quad \forall b \geq 0$$

then

$$|x(t)| \leq e^P |x(t_0)| e^{-\lambda(t-t_0)} + D e^{2P} (t - t_0) e^{-\alpha(t-t_0)} \quad (10)$$

for all $\lambda > 0$ and $\gamma > 0$ where $\alpha = \min\{\lambda, \gamma\}$. From (10) we have that

$$\begin{aligned} |x(t)| &\leq (\rho(t - t_0) + \sigma) e^{-\alpha(t-t_0)} \\ &= (\rho(t - t_0) + \sigma) e^{-(t-t_0)} e^{-(\alpha-\epsilon)(t-t_0)} \end{aligned} \quad (11)$$

where

$$\rho = D e^{2P}, \quad \sigma = e^P |x(t_0)| \quad (12)$$

since $e^{-(\lambda-\alpha)(t-t_0)} \leq 1$ for $t \geq t_0$. By comparison we find that

$$(\rho t + \sigma) e^{-\epsilon t} \leq \xi \quad \forall t \geq 0 \quad (13)$$

if

$$\xi = \begin{cases} \frac{\rho}{\epsilon} e^{\frac{\epsilon^2}{\rho}}, & \rho > \epsilon \sigma \\ \frac{\rho}{\epsilon \sigma}, & \rho \leq \epsilon \sigma. \end{cases}$$

Since

$$e^{\frac{\epsilon^2}{\rho}} < e$$

when $\rho > \epsilon \sigma$, (13) will also be satisfied with the following choice of ξ

$$\xi = \begin{cases} \frac{\rho}{\epsilon}, & \rho > \epsilon \sigma \\ \frac{\rho}{\epsilon \sigma}, & \rho \leq \epsilon \sigma \end{cases} \quad (14)$$

Then, from (11), (12), (13) and (14) we get that

$$\begin{aligned} |x(t)| &\leq \xi e^{-(\alpha-\epsilon)(t-t_0)} \\ &\leq \left(\frac{\rho}{\epsilon} + \sigma \right) e^{-(\alpha-\epsilon)(t-t_0)} \\ &\leq \epsilon (|x(t_0)| + D) e^{-(\alpha-\epsilon)(t-t_0)} \end{aligned}$$

where $\alpha = \min\{\lambda, \gamma\}$ and $\epsilon = \max\{\epsilon^P, \frac{\epsilon^{2P}}{\epsilon}\}$ \square

This lemma implies that a solution $x(t)$ of (4) converges exponentially to zero if $a(x, t)$ and $d(x, t)$ have the properties (5) and (6).

Remark 1. By choosing $\epsilon = \alpha$ we see that

$$\max_{t \geq t_0} |x(t)| \leq (|x(t_0)| + D) \max \left\{ e^P, \frac{e^{2P}}{\alpha} \right\}$$

where $\alpha = \min\{\lambda, \gamma\} > 0$.

Remark 2: If $a(x, t)$ and $d(x, t)$ are continuous in x and t , then there exists at least one solution of (4), [21, Theorem 2.3].

IV. THE SYSTEM

The nonholonomic chained system considered in this paper is the so-called chained form introduced in [26]

$$\begin{aligned}\dot{x}_1 &= u_1 \\ \dot{x}_2 &= u_2 \\ \dot{x}_3 &= x_2 u_1 \\ &\vdots \\ \dot{x}_n &= x_{n-1} u_1.\end{aligned}\quad (15)$$

The input vector fields are

$$\begin{aligned}g_1(x) &= [1, 0, x_2, x_3, \dots, x_{n-1}]^T, \\ g_2(x) &= [0, 1, 0, 0, \dots, 0]^T\end{aligned}$$

where $x = [x_1, x_2, \dots, x_n]^T$. Repeated Lie brackets can be defined recursively by

$$\text{ad}_{g_1}^0 g_2 = g_2, \quad \text{ad}_{g_1}^i g_2 = [g_1, \text{ad}_{g_1}^{i-1} g_2].$$

It can easily be shown that for system (15)

$$\text{ad}_{g_1}^{j-2} g_2 = (-1)^j e_j, \quad j \in \{2, 3, \dots, n\}$$

where $e_j \triangleq [\delta_{1j}, \delta_{2j}, \dots, \delta_{nj}]^T$ where δ_{ij} is the Kronecker delta. Since

$$\forall x \in R^n, \quad \text{span}\{g_1, \text{ad}_{g_1}^0 g_2, \dots, \text{ad}_{g_1}^{n-2} g_2\}(x) = R^n$$

the chained form (15) is completely controllable and the degree of nonholonomy is $n - 1$. In spite of the controllability, (15) cannot be stabilized by a smooth static-state feedback control law as stated by Brockett's Theorem [4]. From [10] we know that to exponentially stabilize (15), the controller cannot be smooth, even if it is time dependent.

V. THE CONTROL LAW

A. Outline of the Control Law and Analysis

In this section, we will propose a control law to globally, \mathcal{K} -exponentially stabilize the nonholonomic chained system (15) about the origin. Since there is no smooth static-state feedback control law which can stabilize (15), we let the feedback control law be time dependent, as first proposed by [32] for a nonholonomic cart. The main idea of the control scheme to be presented is based on the observation that if u_1 is a function of time only, then the state variables

$$z \triangleq [x_2, \dots, x_n]^T \quad (16)$$

satisfy a linear time-varying state space model with input u_2 . This linear time-varying system with input u_2 and state z can be made \mathcal{K} -exponentially stable using feedback from z , which will be shown in the following. The problem which remains to be solved is how to achieve \mathcal{K} -exponential stability when x_1 is included into the state vector, that is, when the complete system with state vector x and inputs u_1 and u_2 is considered. The solution proposed in this paper is as follows: Define

a sequence (t_0, t_1, t_2, \dots) where the equidistant terms are defined by

$$t_i \triangleq iT \quad (17)$$

where T is a strictly positive constant. Then, the control u_1 is given by $u_1 = u_1(t, x(t_i))$, $t \in [t_i, t_{i+1})$, that is, u_1 is a function of time and not of the state whenever $t_i < t < t_{i+1}$. By appropriately selecting $u_1(t, x(t_i))$, it is then possible to achieve exponential convergence in $x_1(t)$ and, in addition, by patching together the solutions for $z(t)$, \mathcal{K} -exponential stability for the complete system can be established. To obtain exponential convergence, the time-varying feedback control law is nonsmooth in the origin with respect to the state $x(t_i)$.

In this section the control laws are presented without further explanation. Then, in the following sections the stability of the closed-loop system is analyzed, and it is shown that the proposed control laws have the desired properties. The analysis is done in two steps: First the stability of the linear system with state vector z and input u_2 is analyzed. To this end a vector $z^d(x, t): R^n \times R_+ \rightarrow R^{n-1}$ is defined, and it is shown that for the proposed control law the vector $\tilde{z}(x, t) \triangleq z - z^d(x, t)$ converges exponentially to zero. Then, this result is used to prove exponential convergence of $z(t)$. At this point the state x_1 is included, and \mathcal{K} -exponential stability of the complete system with state vector x is established.

Concerning the notation, throughout the paper the one-norm will be used, i.e., the norm of an n dimensional vector $x = [x_1, \dots, x_n]^T$ is given by

$$\|x\| \triangleq \sum_{j=1}^n |x_j|. \quad (18)$$

Furthermore, the index i will refer to terms in the time sequence (t_0, t_1, \dots) .

B. The Control Law for u_1

Let $k(\cdot)$ be a bounded function

$$k: R^n \rightarrow R \quad (19)$$

which satisfies for a strictly positive constant K

$$\forall x \in R^n |k(x)| \leq K, \quad \forall x \in R^n / \{0\} k(x) \neq 0, \quad k(0) = 0. \quad (20)$$

The idea is now to let u_1 be given by a time-varying function multiplied with the function $k(x(t_i))$ such that the part of the system which is represented by the state variables z with input u_2 is linear and time-varying in the time interval (t_i, t_{i+1}) . The input u_2 is then used to make $[x_2(t), \dots, x_n(t)]^T$ exponentially converge to zero. The sign and magnitude of $k(x(t_i))$ will be chosen such that $x_1(t)$ also converges exponentially to zero, and the inputs remain bounded.

We introduce a time-varying function $f: R_+ \rightarrow R$ which has the following properties:

- P1) $f(t) \in C^\infty[t_0, +\infty)$
- P2) $0 \leq f(t) \leq 1, \forall t \geq t_0$

P3) $f(t_i) = 0$, $t_i \in \{t_0, t_1, t_2, \dots\}$

P4) for any $j \in \{3, \dots, n\}$, there are positive constants η_j and P_j such that

$$\forall t_p \in \{t_0, t_1, \dots\}, \forall t \geq t_p \quad \int_{t_p}^t [f^{2j-3}(\tau) - \eta_j] d\tau \leq P_j.$$

A function satisfying conditions P1)–P4) is

$$f(t) = (1 - \cos \omega t)/2, \quad \omega = \frac{2\pi}{T} \quad (21)$$

where $T = t_{i+1} - t_i$ is a constant, (17). Incidentally, the function $f(t)$ is not restricted to be time-periodic to satisfy P1)–P4).

The control law for u_1 is defined by

$$u_1 = k(x(t_i))f(t), \quad t \in [t_i, t_{i+1}) \quad (22)$$

where

$$k(x) = \text{sat}(-[x_1 + \text{sgn}(x_1)G(\|z\|)]\beta, K). \quad (23)$$

Here

$$\text{sat}(q, k) \triangleq \begin{cases} q & |q| < K \\ K \text{sgn}(q) & |q| \geq K \end{cases}$$

and

$$G(\|z\|) = \kappa \|z\|^{\frac{1}{2n-1}} \quad (24)$$

$$\beta = \frac{1}{\int_{t_i}^{t_{i+1}} f(\tau) d\tau} \quad (25)$$

where κ is a positive constant and $\text{sgn}(x_1)$ is defined as

$$\text{sgn}(x_1) = \begin{cases} 1, & x_1 \geq 0 \\ -1, & x_1 < 0 \end{cases}.$$

All the norms are one-norms as defined in (18). The saturation function is used to guarantee global stability. The idea of using saturation functions was also used in [34].

Note that due to P1)–P3) and the bound on $k(\cdot)$, the control u_1 is bounded and continuous with respect to time.

C. The Control Law for u_2

With the input u_1 given by (22), we get from (15) and (16) that \dot{z} is given by

$$\dot{x}_2 = u_2$$

$$\dot{x}_3 = k(x(t_i))f(t)x_2$$

$$\dot{x}_n = k(x(t_i))f(t)x_{n-1}. \quad (26)$$

In this subsection, we derive a feedback control law for u_2 to make $z(t) = [x_2(t), \dots, x_n(t)]^T$ globally, exponentially converge to zero.

Define $g_{jm}: R_+ \rightarrow R$ for $j, m \in \{2, \dots, n\}$ by

$$g_{n-1,n} = -\lambda_n \quad (27)$$

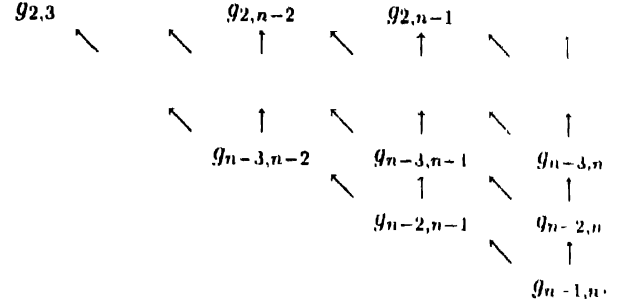
$$g_{j-1,m}(t) = -g_{jm}[\lambda_j f^{2j-2}(t) + 2(j-1)\dot{f}(t) + f(t)(\dot{g}_{jm}(t) + g_{j,m+1}(t)f(t))] \quad (28)$$

$$g_{j-1,j}(t) = -\lambda_j + f^2(t)g_{j,j+1}(t) \quad (29)$$

$$g_{jp} = 0 \quad \text{if } p \leq j \text{ or } p = n+1 \quad (30)$$

where λ_j , $j \in \{2, \dots, n\}$, are any positive constants. Incidentally, we see from this definition of $g_{jm}(t)$ and P1)–P4) that $g_{jm}(t)$ is smooth.

The definition of $g_{jm}(t)$ from (27)–(30) can be illustrated by the following diagram where $a \rightarrow b$ means that b depends on a .



The control law for u_2 is defined for $t \in [t_i, t_{i+1})$ by

$$u_2 = \begin{cases} \Gamma(k(x(t_i)), t)^T z, & x(t_i) \neq 0 \\ 0, & x(t_i) = 0 \end{cases} \quad (31)$$

where

$$z = [x_2, \dots, x_n]^T \in R^{n-1},$$

$$\Gamma(k, t) = [\Gamma_2(k, t), \dots, \Gamma_n(k, t)]^T \in R^{n-1}$$

and for $j \in \{3, \dots, n\}$

$$\Gamma_2(k, t) = -\lambda_2 + f^3 g_{2,3}$$

$$\Gamma_j(k, t) = f(\lambda_2 f g_{2j} + 2\dot{f} g_{2j} + f\dot{g}_{2j} + f^2 g_{2,j+1}) \frac{1}{k^{j-2}}.$$

The smooth functions $g_{2j}(t)$ are given by (27)–(30). The function $k(x(t_i))$ is given by (23).

The feedback control law (22) and (31) depends on $k(\cdot)$ which is a nonsmooth function of the state $x(t_i)$. One might, therefore, raise the question about the existence and uniqueness of the solution $x(t)$ of (15). Since in each time interval $[t_i, t_{i+1})$ the closed-loop system (15) with the control law (22) and (31) becomes a linear, time-varying system, global existence and uniqueness of $x(t)$ can be shown in any time interval $[t_i, t_{i+1})$ for any $x(t_i)$ by applying e.g., Theorem 2.3 in [14]. The time intervals can then be patched together to show existence and uniqueness of $x(t)$ for all $t \in R_+$ and for all initial conditions $x(t_0)$.

VI. CONVERGENCE ANALYSIS OF $z(t)$

In this section, it is shown that the control law (22) and (31) makes $z(t)$ (16) globally exponentially converge to zero.

To this end, define the functions $z^d = [x_2^d, \dots, x_n^d]^T: R_+ \times R_+ \rightarrow R^{n-1}$ as follows for $t \in [t_i, t_{i+1})$

$$x_n^d \triangleq 0 \quad (32)$$

$$x_j^d(x, t) \triangleq \begin{cases} f^{2j-2}(t) \sum_{m=j+1}^n g_{jm}(t) \frac{1}{k^{m-j}(x(t_i))} x_m, & x(t_i) \neq 0 \\ 0, & x(t_i) = 0 \end{cases} \quad (33)$$

Note from this definition of x_j^d that the control law for u_2 (31) satisfies

$$u_2 = -\lambda_2(x_2 - x_2^d) + \dot{x}_2^d \quad (34)$$

whenever $x(t_i) \neq 0$. This relation will be used in the convergence analysis.

By assuming that the control laws yield a continuous solution $x(t)$ of (15), we see from the definition (33) of $x_j^d(x, t)$, P1), and P3) that $x_j^d(x(t), t)$ is continuous for all $t > 0$ and $x_j^d(x(t_i), t_i) = 0$. In addition, z^d has the properties shown in the following lemma.

Lemma 2: Let $z^d = [x_2^d, \dots, x_n^d]^T$ be defined by (32) and (33). Then, for $j \in \{3, \dots, n\}$, $t \in [t_i, t_{i+1})$

$$\dot{x}_n^d = 0 \quad (35)$$

$$x_{j-1}^d k(x(t_i)) f(t) = -\lambda_j f^{2j-3}(t) [x_j - x_j^d(x, t)] + \dot{x}_j^d(x, t). \quad (36)$$

Proof: Equation (35) is trivially satisfied from (32). Equation (36) will be proved by calculating $\dot{x}_j^d(x, t)$. The arguments of the functions $f(t)$, $k(x(t_i))$, and $x_j^d(x, t)$ will be omitted in the proof for simplicity. From the definition of x_j^d , (33), we find for $t \in (t_i, t_{i+1})$ in the case $x(t_i) \neq 0$ by using $\dot{x}_m = x_{m-1} k f$ from (15)

$$\begin{aligned} \dot{x}_j^d &= (2j-2)f^{2j-3} \dot{f} \sum_{m=j+1}^n g_{jm} \frac{1}{k^{m-j}} x_m \\ &+ f^{2j-2} \sum_{m=j+1}^n \left[\dot{g}_{jm} \frac{1}{k^{m-j}} x_m + g_{jm} \frac{f}{k^{m-j-1}} x_{m-1} \right]. \end{aligned} \quad (37)$$

We see that by assuming a continuous solution $x(t)$ of (15), we have due to P1) and P3)

$$\lim_{t \rightarrow t_i^+} \dot{x}_j^d(x(t), t) = \lim_{t \rightarrow t_i} \dot{x}_j^d(x(t), t) = 0.$$

Therefore, (37) is valid for all $t > 0$. Now, we will express x_{j-1}^d by using \dot{x}_j^d from (37). From (33), (28)–(30) we find

$$\begin{aligned} x_{j-1}^d &= f^{2j-4} \sum_{m=j}^n g_{j-1,m} \frac{1}{k^{m-(j-1)}} x_m \\ &= f^{2j-4} \left\{ x_j \left(-\frac{\lambda_j}{k} + f^2 g_{j,j+1} \frac{1}{k} \right) \right. \\ &\quad + \sum_{m=j+1}^n \frac{x_m}{k^{m-j+1}} [(\lambda_j f^{2j-2} + 2(j-1)\dot{f}) g_{jm} \\ &\quad \left. + f(\dot{g}_{jm} + g_{j,m+1} f)] \right\} \\ &= -\frac{\lambda_j}{k} f^{2j-4} \left(x_j - f^{2j-2} \sum_{m=j+1}^n g_{jm} \frac{1}{k^{m-j}} x_m \right) \\ &\quad + f^{2j-4} \left\{ (2j-2)\dot{f} \sum_{m=j+1}^n g_{jm} \frac{1}{k^{m-j+1}} x_m \right. \\ &\quad \left. + f \sum_{m=j+1}^n \left[\dot{g}_{jm} \frac{1}{k^{m-j+1}} x_m + g_{jm} \frac{f}{k^{m-j}} x_{m-1} \right] \right\} \\ &= -\frac{\lambda_j}{k} f^{2j-4} (x_j - x_j^d) + \frac{1}{kf} \dot{x}_j^d. \end{aligned} \quad (38)$$

Equation (36) follows readily. \square

Define the functions $\tilde{z} = [\tilde{x}_2, \dots, \tilde{x}_n]^T: R^n \times R_+ \rightarrow R^{n-1}$ by

$$\tilde{x}_j(x, t) \triangleq x_j - x_j^d(x, t), \quad j \in \{2, \dots, n\}. \quad (39)$$

From the model (15), Lemma 2, and the control law (22) and (31) with the expression (34) for the input u_2 , we find that \tilde{z} satisfies

$$\dot{\tilde{x}}_2 = -\lambda_2 \tilde{x}_2 \quad (40)$$

$$\dot{\tilde{x}}_3 = \lambda_3 f^3(t) \tilde{x}_3 + k f(t) \tilde{x}_2$$

$$\vdots$$

$$\dot{\tilde{x}}_n = -\lambda_n f^{2(n-2)+1}(t) \tilde{x}_n + k f(t) \tilde{x}_{n-1}. \quad (41)$$

The following lemma shows that $\tilde{x}_j(x(t), t)$ tends exponentially toward zero.

We denote

$$\underline{z}_j = [x_2, x_3, \dots, x_j]^T, \quad j \in \{2, \dots, n\} \quad (42)$$

and use the one-norm $\|\cdot\|$ as defined by (18).

Lemma 3: Consider system (15) with the control law (22) and (31). Let $\tilde{z}(x(t), t)$ be defined by (39) where $x(t)$ is the solution of (15). Then, for any $j \in \{2, \dots, n\}$, there is a $\delta > 0$ such that

$$\forall \varepsilon_2, \quad \varepsilon_j \in (0, \delta),$$

$$\exists c_j > 0, \exists \gamma_j > 0,$$

$$\forall t_p \in \{t_0, t_1, \dots\}, \forall \underline{z}_j(t_p) \in R^{j-1}$$

$$|\tilde{x}(x(t), t)| \leq c_j \|\underline{z}_j(t_p)\| e^{-\gamma_j(t-t_p)}, \quad \forall t \geq t_p \quad (43)$$

where the constants γ_j can be defined as

$$\gamma_j = \alpha_j - \varepsilon_j > 0 \quad (44)$$

$$\alpha_q = \min\{\lambda_q \eta_q, \alpha_{q-1} - \varepsilon_{q-1}\}, \quad q \in \{3, \dots, j\} \quad (45)$$

$$\alpha_2 = \lambda_2. \quad (46)$$

The constants η_q are found from Property P4) of $f(t)$.

Proof The proof will be given by induction. Assume that (43) is satisfied for $j = m-1 \in \{2, \dots, n-1\}$ for a $t_p \in \{t_0, t_1, \dots\}$. Then, from (40)–(41) we have that

$$\begin{aligned} \dot{\tilde{x}} &= -\lambda_m f^{2m-3}(t) \tilde{x}_m + k(x(t_i)) f(t) \tilde{x}_{m-1} \\ &= -a(t) \tilde{x}_m + d(t) \end{aligned}$$

where

$$a(t) = \lambda_m f^{2m-3}(t)$$

$$d(t) = k(x(t_i)) f(t) \tilde{x}_{m-1}(x(t), t).$$

From Property P4) of $f(t)$ we have that for all $t_p \in \{t_0, t_1, \dots\}$ and for all $t \geq t_p$

$$\int_{t_p}^t (a(\tau) - \lambda_m \eta_m) d\tau \leq \lambda_m P_m.$$

Since (43) is assumed to be satisfied for $j = m-1$ and since $|f(t)| \leq 1$ from P1), and $|k(x)| \leq K$ from (20) and (23) we have

$$|d_m(t)| \leq D_m e^{-\gamma_{m-1}(t-t_p)}$$

here $D_m \triangleq K c_{m-1} \|\bar{z}_{m-1}(t_p)\|$. From the definition of $x_j(t)$ (39) and since $x_j^d(x(t_p), t_p) = 0$ for all $t_p \in \{0, t_1, \dots\}$, Section V-C, we have $\bar{x}_m(x(t_p), t_p) = x_m(t_p) = \bar{z}_m(t_p)$. From Lemma 1 we can then conclude that

$$\begin{aligned} \forall \varepsilon_m > 0 \exists c_m \exists \gamma_m \forall \bar{z}_{m-1} | \\ |\bar{x}_m(t)| &\leq c_m (\|\bar{z}_{m-1}(t_p)\| + |\bar{x}_m(t_p)|) e^{-\gamma_m(t-t_p)} \\ &= c_m \|\bar{z}_{m-1}(t_p)\| e^{-\gamma_m(t-t_p)} \end{aligned}$$

where the constants γ_m and c_m can be defined as

$$\begin{aligned} \gamma_m &= \alpha_m - \varepsilon_m \\ \alpha_m &= \min\{\lambda_m \eta_m, \gamma_{m-1}\} \\ c_m &= \max\left\{e^{\lambda_m P_m}, e^{2\lambda_m P_m} \frac{K c_{m-1}}{\varepsilon_m}\right\}. \end{aligned}$$

The constant γ_m is strictly positive if

$$0 < \varepsilon_j < \alpha_j, \quad j \in \{2, \dots, m\}.$$

Consequently, there exists a constant δ such that $\varepsilon_2, \dots, \varepsilon_n \in (0, \delta)$ implies that $\gamma_m > 0$.

We have, therefore, proved that if (43) is satisfied for $j = m-1$ for a $t_p \in \{t_0, t_1, \dots\}$, then (43) is satisfied for $j = m$, for all $m \in \{3, \dots, n\}$. The induction is completed by showing that (43) is satisfied for $j = 2$. From (40) we find that for all $x(t_p), t_p \in \{t_0, t_1, \dots\}$

$$\begin{aligned} |\bar{x}_2(x(t), t)| &= |\bar{x}_2(x(t_p), t_p)| e^{-\lambda_2(t-t_p)} \\ &= \|\bar{z}_2(t_p)\| e^{-\lambda_2(t-t_p)} \end{aligned}$$

since $x_j^d(x(t_p), t_p) = 0$ for $t_p \in \{t_0, t_1, \dots\}$ and hence $x_j(x(t_p), t_p) = x_j(t_p)$. Equation (43) is then satisfied for $j = 2$ by choosing $\varepsilon_2 = 1$. (In this case, (43) is also satisfied for $\varepsilon_2 = 0$.) Since (43) is satisfied for all $t_p \in \{t_0, t_1, \dots\}$ when $j = 2$ then (43) is proved by induction to be satisfied for all $t_p \in \{t_0, t_1, \dots\}$. \square

Remark: Note from (45) that arbitrarily large values for $\alpha_2, \dots, \alpha_n$ can be obtained by choosing $\lambda_2, \dots, \lambda_n$ appropriately, which means that the exponential rate of convergence $\gamma_j = \alpha_j - \varepsilon_j$ can be chosen arbitrarily large since ε_j can be chosen arbitrarily close to zero.

The definitions of x^d (32)–(33) and \hat{x} (39) can be used to show that the state x_j can be expressed as a weighted sum of the functions $\tilde{x}_r, r \in \{j, \dots, n\}$ as stated in the following lemma.

Lemma 4: Let $x_j^d(x, t)$ be defined by (32)–(33) and $\hat{x}_j(x, t)$ by (39) for $j \in \{2, \dots, n\}$ and $t \in [t_i, t_{i+1})$. If $x(t_i) \neq 0$ then x_j satisfies for all $j \in \{2, \dots, n-1\}$

$$x_j = \tilde{x}_j(x, t) + f^{2j-2}(t) \sum_{r=j+1}^n \tilde{g}_{jr}(t) \frac{1}{k^{r-j}(x(t_i))} \hat{x}_r(x, t). \quad (47)$$

The functions $\tilde{g}_{jr}: R_+ \rightarrow R$ are bounded and smooth.

Proof: See Appendix A. \square

From Lemmas 3 and 4 we can now show global, exponential convergence of $z(t)$ to zero.

Theorem 1: Consider system (15) with the control (22) and (31) and z defined in (16). There are a class \mathcal{K} function $h_z: R_+ \rightarrow R_+$ of class \mathcal{K} and a constant $\gamma_z > 0$ such that, for any initial condition $z(t_0) \in R^{n-1}$

$$\|z(t)\| \leq h_z(\|z(t_0)\|) e^{-\gamma_z(t-t_0)}, \quad \forall t \geq t_0$$

The constant γ_z can be given by

where γ_n is given from (44). \square

Proof: See Appendix A. \square

VII. STABILITY ANALYSIS OF $x(t)$

In this section, we prove global \mathcal{K} -exponential stability of the state $x(t)$ of the system (15) about zero.

By showing that $u_1 = k(x(t))f(t)$, (22), makes $x_1(t)$ converge to zero as $z(t)$ converge to zero, Theorem 1 can be used to show the following lemma.

Lemma 5: Let the control law be given by (22) and (31). Then, system (15) is \mathcal{K} -exponentially stable about the origin, i.e., in a neighborhood Ω about $x = 0$ there exist a function $h(\cdot; T)$ of class \mathcal{K} and a constant $\gamma > 0$ such that

$$\forall x(t_0) \in \Omega \|x(t)\| \leq h(\|x(t_0); T\|) e^{-\gamma(t-t_0)}, \quad \forall t \geq t_0 \quad (49)$$

where the function $h(\cdot; T)$ depends on time parameter $T = t_{i+1} - t_i$, (17). \square

Proof: See Appendix A. \square

The neighborhood Ω is given by

$$\Omega = \left\{x \mid |x_1| < \frac{K}{2\beta}, \quad G(h_2(\|z\|)) < \frac{K}{2\beta}\right\} \quad (50)$$

where $h_z(\cdot)$ is a function of class \mathcal{K} from Theorem 1 and $z = [x_2, \dots, x_n]^T$ (16). The function $G(\cdot)$ is defined in (24), and β is defined in (25). The constant γ can be defined as

$$\begin{aligned} \gamma &= \frac{\gamma_z}{2n-4} > 0 \\ \gamma_z &= \frac{\gamma_n}{2} \end{aligned}$$

where γ_n is given from Lemma 3.

Remark 1: The exponential convergence rate γ in (49) can be selected arbitrarily large by choosing $\lambda_2, \dots, \lambda_n$ appropriately, that is large enough. See also the remark to Lemma 3 in Section VI. For all $\lambda_2, \dots, \lambda_n > 0$ the system is \mathcal{K} -exponentially stable for all positive T where $T = t_{i+1} - t_i$. The class \mathcal{K} function $h(\cdot; T)$ increases exponentially, however, with the parameter T . This is due to the fact that it takes at least the time T for the input u_1 to drive x_1 to an arbitrarily small neighborhood about zero.

Remark 2. The neighborhood Ω can be chosen arbitrarily large by choosing the parameters K , \mathcal{K} and β (or $T = t_{i+1} - t_i$) appropriately. Therefore, the stability is semi-global.

Remark 3: The bounds on $\|x(t)\|$ given by the function $h(\cdot; T)$, (88), and γ_s can be very conservative and do not in general provide quantitative information.

Lemma 5 does not prove global, \mathcal{K} -exponential stability because of the saturation function in the definition of $k(x(t))$, (23). A function $h_x(\cdot)$ of class \mathcal{K} , however, can be constructed to yield also global, \mathcal{K} -exponential stability as shown in the following theorem.

Theorem 2: Let the control law be given by (22) and (31). Then system (15) is globally, \mathcal{K} -exponentially stable about the origin, i.e., there exist a function $h_x(\cdot; T)$ of class \mathcal{K} and a constant $\gamma > 0$ such that $\forall t \geq t_0$

$$\forall x(t_0) \in R^n \|x(t)\| \leq h_x(\|x(t_0)\|; T e^{-\gamma(t-t_0)}). \quad (51)$$

Proof: See Appendix A. \square

As for Lemma 5, the constant γ can be defined as follows

$$\gamma = \frac{\gamma_s}{2n-1} > 0$$

$$\gamma_s = \frac{\gamma_n}{2}$$

where γ_n is given from Lemma 3.

The remarks after Lemma 5 on the rate of convergence γ and the dependence of T for the class \mathcal{K} function $h_x(\cdot; T)$ also apply here.

VIII. STABILIZATION ABOUT ARBITRARY CONFIGURATION

We have shown that the controller (22) and (31), and (23) makes system (15) globally, \mathcal{K} -exponentially stable about the origin, i.e., about $[x_1, \dots, x_n]^T = 0$. In this section we show that the same control law can be used to \mathcal{K} -exponentially stabilize the chained form (15) about any configuration. Indeed, here we show that any strategy to control the chained form to the origin can be used to control it to any desired configuration.

Let the desired constant configuration be given by

$$x^p = [x_1^p, x_2^p, \dots, x_n^p]^T \in R^n, \quad \dot{x}^p \equiv 0.$$

Now we introduce the following variables

$$x_m^r \triangleq x_m^p + \sum_{j=2}^{m-1} x_j^p \frac{1}{(m-j)!} (x_1 - x_1^p)^{m-j}, \quad m \in \{1, \dots, n\} \quad (52)$$

Here, x_i is a state variable of the chained form (15) satisfying $\dot{x}_1 = u_1$. The vector $x^r = [x_1^r, \dots, x_n^r]^T$ is thus given from (52) as a smooth function $x^r = \phi(x_1; x^p)$.

Lemma 6: Let $x^r = [x_1^r, \dots, x_n^r]^T$ be given by (52). Then

$$\dot{x}_1^r = 0 \quad (53)$$

$$\dot{x}_2^r = 0 \quad (54)$$

$$\dot{x}_m^r = x_{m-1}^r u_1, \quad m \in \{3, \dots, n\}. \quad (55)$$

Proof: From (52) we have that $x_1^r = x_1^p$ and $x_2^r = x_2^p$. Since x_1^p and x_2^p are constants, (53)–(54) follow readily. Equation (55) can be proved by induction. Assume that there is an index $m \in \{3, \dots, n-1\}$ such that

$$\dot{x}_m^r = x_{m-1}^r u_1.$$

Since $\dot{x}_m^p = 0$, $m \in \{1, \dots, n\}$, differentiating x_{m+1}^r from (52) gives

$$\begin{aligned} \dot{x}_{m+1}^r &= \sum_{j=2}^m x_j^p \frac{1}{(m+1-j)!} (m+1-j) \\ &\quad \times (x_1 - x_1^p)^{m-j} (\dot{x}_1 - \dot{x}_1^p) \\ &= \sum_{j=2}^m x_j^p \frac{1}{(m-j)!} (x_1 - x_1^p)^{m-j} u_1 \\ &= \left[x_m^p + \sum_{j=2}^{m-1} x_j^p \frac{1}{(m-j)!} (x_1 - x_1^p)^{m-j} \right] u_1 \\ &= x_m^r u_1. \end{aligned}$$

Consequently, if (55) is satisfied for index m then (55) is satisfied for $m+1$, too. The proof is then completed by showing that $\dot{x}_1^r = x_2^r u_1$. From (52) we get

$$\dot{x}_1^r = \frac{d}{dt} [x_2^p (x_1 - x_1^p) + x_1^p] = x_2^p u_1 = x_2^r u_1.$$

\square

We now define

$$\bar{x}_m \triangleq x_m - x_m^p, \quad m \in \{1, \dots, n\}. \quad (56)$$

From system (15) and Lemma 6 it follows that

$$\dot{\bar{x}}_1 = u_1 \quad (57)$$

$$\dot{\bar{x}}_2 = u_2 \quad (58)$$

$$\dot{\bar{x}}_m = \bar{x}_{m-1} u_1, \quad m \in \{3, \dots, n\}. \quad (59)$$

This system has the same structure as the chained form (15). A control law for (15) controlling $x = [x_1, \dots, x_n]^T$ to zero can be used to control $\bar{x} = [\bar{x}_1, \dots, \bar{x}_n]^T$ to zero. The coordinate transformation between x and \bar{x} is then given from (52) and (56) by

$$\bar{x} = x - x^p = x - \phi(x_1; x^p) \triangleq \tau(x; x^p)$$

$$x = \bar{x} + x^p = \bar{x} + \phi(x_1; x^p)$$

$$= \bar{x} + \phi(x_1 + x_1^p; x^p) \triangleq \bar{\tau}(\bar{x}; x^p)$$

where $\tau(\cdot; x^p)$ and $\bar{\tau}(\cdot; x^p)$ are smooth functions.

We can then conclude with the following theorem.

Theorem 3: Given the system (57)–(59) where \bar{x}_m and x_m^r are given by (56) and (52). Then, a control law for (57)–(59) making $\bar{x} = [\bar{x}_1, \dots, \bar{x}_n]^T$ converge to zero makes x converge to the desired configuration x^p . The convergence of x to x^p is exponential if \bar{x} converges exponentially to zero.

Proof. From (56) and (52) we have that

$$\begin{aligned} x_m &= x_m^r + \bar{x}_m = x_m^p + \sum_{j=2}^{m-1} x_j^p \frac{1}{(m-j)!} \bar{x}_1^{m-j} \\ &\quad + \bar{x}_m, \quad m \in \{1, \dots, n\}. \end{aligned}$$

Therefore, the control of \bar{x} to zero implies the control of x to x^p . We also see that exponential convergence of \bar{x} to zero implies exponential convergence of x to x^p . \square

IX. SIMULATIONS

A simulation with $n = 4$ was done in MATLAB at a SPARC station 1. The control law for u_1 was chosen as follows

$$u_1 = k(x(t_i))f(t), \quad f(t) = (1 - \cos t)/2$$

where

$$k = \text{sat}(-[x_1(t_i) + \text{sgn}(x_1(t_i))G(\|z(t_i)\|)]\beta, K), \quad K = 2$$

as defined in (23). By studying the time integral of $f^3(t)$ and $f^5(t)$, we see that Property P4) of $f(t)$ is satisfied by choosing

$$\eta_3 = \frac{5}{16}, \quad \eta_4 = \frac{63}{256}, \quad P_3 = \frac{1}{2}, \quad P_4 = \frac{1}{2}.$$

The controller parameter κ in $G(\cdot)$, (24), was taken to $\kappa = 3$. The constant β is given by (25)

$$\beta = 1 / \int_0^T f(\tau) d\tau = \frac{1}{\pi}$$

where T is the time-period of the function $f(t)$, i.e., $T = 2\pi$. The instants of time t_i where $k(x(t_i))$ may switch are given by the set $\{0, 2\pi, 4\pi, 6\pi, \dots\}$.

We find the control law for u_2 from (31) and (27)–(30)

$$\begin{aligned} u_2 = & -(\lambda(1 + f^3 + f^5))x_2 \\ & - (f\lambda + f^3\lambda + f^5\lambda + 2\dot{f} + 8f^2\dot{f})/k x_3 \\ & - (f\lambda(f^5\lambda^2 + 4f\lambda\dot{f} + 6f^4\lambda\dot{f} + 8\dot{f}^2 + 4f\ddot{f})/k^2)x_4. \end{aligned}$$

Here, we have chosen

$$\lambda = \lambda_2 = \lambda_3 = \lambda_4.$$

In the simulation, λ was taken to $\lambda = 1$.

The initial state was chosen as

$$(x_1(0), x_2(0), x_3(0), x_4(0)) = (0, -0.1, 0.1, 1).$$

Euler's method was applied for the numerical integration where the time-step was taken to 0.05. In Fig. 1, $z(t) = [x_2(t), x_3(t), x_4(t)]^T$ is plotted versus time showing the convergence to zero. The state variable $x_1(t)$ is plotted in Fig. 2. We see that $x_1(t)$ converges to zero, too. Note, however, from the time-axes that the rate of convergence of $x_1(t)$ is slower than the one of $z(t)$. This coincides with the relation between γ and γ_z in Lemma 5 where $\gamma = \gamma_z/4$. To show that the convergence is exponential, $\log \|x(t)\|$ is plotted in Fig. 3 where the one-norm is chosen. We see from this figure that $\log \|x(t)\|$ decreases toward $-\infty$ implying that $\|x(t)\|$ exponentially converges to zero.

By using the coordinate transformation from [33], we can interpret the variables x_1 and x_4 as the x - and y -position of the midpoint of the rear axle of a four-wheeled car. The path $(x(t), y(t))$ is presented in Fig. 4. We see that the motion seems natural when interpreted as a parking maneuver.

In Figs. 5 and 6 the inputs $u_1(t)$ and $u_2(t)$ are shown as functions of time. We see that they are continuous and converge exponentially to zero.

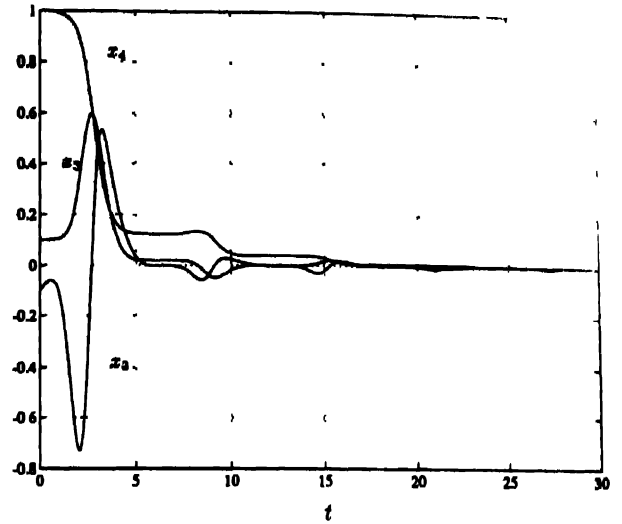


Fig. 1. Exponential convergence of $z(t) = [x_2, x_3, x_4]^T$ to zero

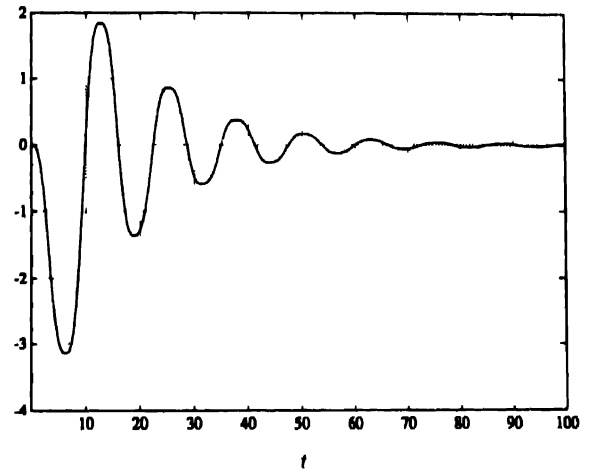


Fig. 2. Exponential convergence of $x_1(t)$ to zero

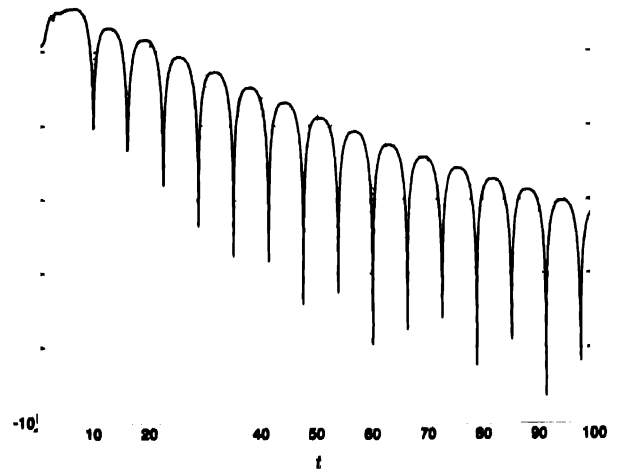


Fig. 3. Convergence of $\log \|x(t)\|$ to $-\infty$.

X. SUMMARY AND CONCLUSION

A feedback control law has been proposed to globally stabilize a chained nonholonomic system of any dimension.

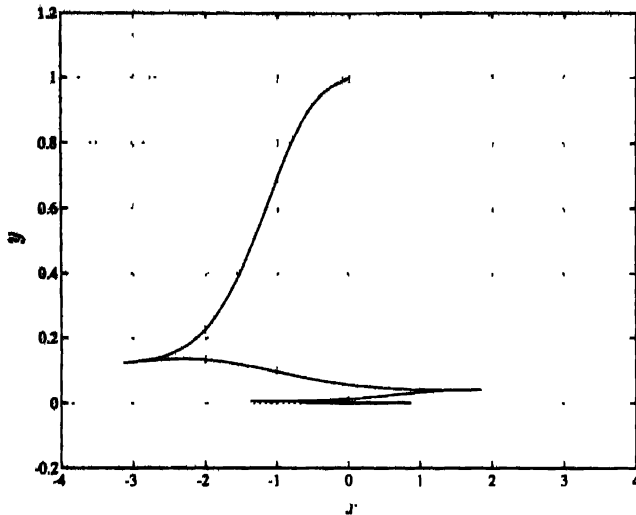


Fig. 4. The resulting path in the xy -plane when applying the exponentially convergent control law. The variables $x = x_1$ and $y = x_4$ are interpreted as the planar position of a four-wheeled car

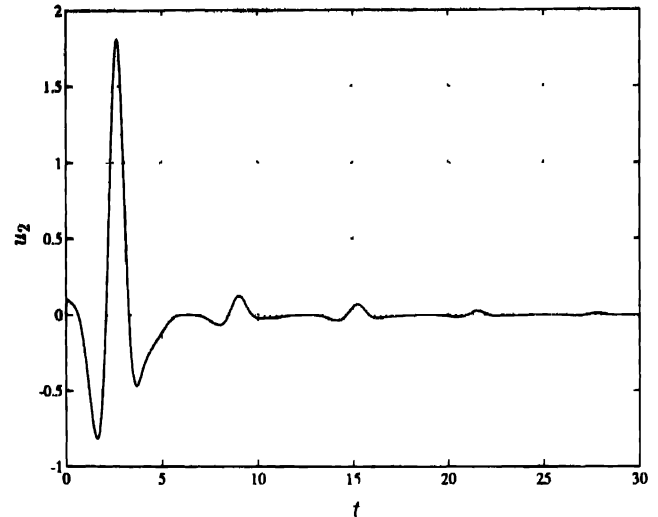


Fig. 6. The input $u_2(t)$ from the exponentially convergent control law

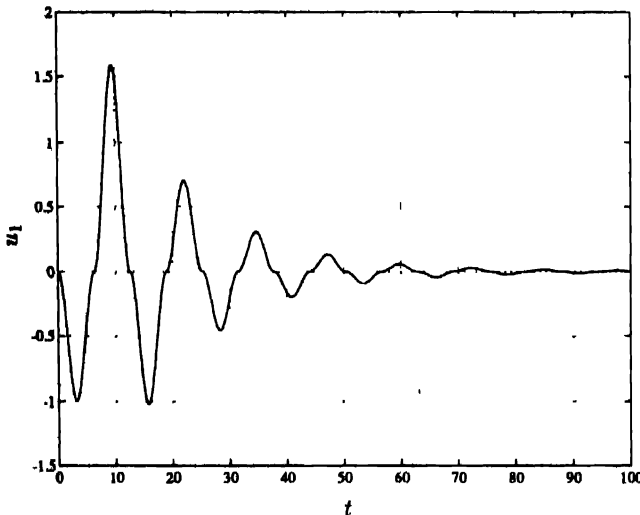


Fig. 5. The input $u_1(t)$ from the exponentially convergent control law

The resulting rate of convergence is exponential and global, \mathcal{K} -exponential stability was achieved. \mathcal{K} -exponential stability is a weaker form of stability than exponential stability in the usual sense, but it possesses the same rate of convergence.

The feedback control law depends on a time-varying function, not necessarily periodic, and on a function which is nonsmooth with respect to the state at predefined discrete instants of time. The exponential rate of convergence can be arbitrarily fast by choosing the controller parameters appropriately. The controller cannot, however, drive the system to an arbitrary small neighborhood of the desired configuration in shorter time than T i.e., the length of the time intervals used in the control law. The class \mathcal{K} function which defines a initial-state-dependent upper bound of the time-evolution of the system, has T as a parameter. Simulation of a four dimensional chained form indicated the exponential convergence. By using a coordinate transformation similar to the transformation from chained form to power form, a strategy to control the state to zero can be used to control the state of any desired configuration.

The idea of introducing a function which varies with the state at discrete instants of time seems, therefore, to be a useful approach to obtain good stability properties for nonholonomic systems. The control law has a simple structure, though the stability analysis is quite involved because of few existing mathematical tools for such systems.

A drawback of this approach is the possibility for numerical problems at digital computers as the parameter $k(x(t_i))$ approaches zero, since there are divisions by $k(x(t_i))$ in the control law for u_2 . (These divisions do not cause unbounded quantities, since $k(x(t_i))$ always dominates the numerator.) Such numerical problems were, however, not observed during the simulations. Since the parameter $k(x(t_i))$ only changes at $t_i \in \{t_0, t_1, \dots\}$, the control of x_1 may be more sensitive to disturbances and modeling imperfections than if $k = k(x(t))$ i.e., function of the state at all $t \geq t_0$. Depending on the physical application, feedback from the state only at $t_i \in \{t_0, t_1, \dots\}$ may be sufficient to make x_1 exponentially converge to zero. Note that the state variables $z = [x_2, \dots, x_n]^T$ are controlled by the control law $u_2 = u_2(z(t), k)$ which depends on $z(t)$ for all $t \geq t_0$.

In further work, an extension from $k = k(x(t_i))$ to $k = k(x(t))$ by redefining z^d , (32)–(33), can be studied. The structure of $k(x(t_i))$ in the present control law indicates that the use of $k = k(x(t))$ results in a nonsmooth, time-varying feedback control law. Also, more mathematical tools are needed to analyze such systems. The need for the definition of \mathcal{K} -exponential stability for the stabilization with exponential convergence of chained systems and other nonholonomic systems should also be analyzed.

APPENDIX PROOFS

Proof of Lemma 4: The arguments of $k(x(t_i))$, $\tilde{x}(x, t)$, $x^d(x, t)$, $f(t)$, and $g_{j,r}(t)$ will mostly be omitted for simplicity. From the definition (39) of \tilde{x}_{n-1} we have that

$$x_{n-1} = \tilde{x}_{n-1} + x_{n-1}^d. \quad (60)$$

om (32) we have that $x_n^d = 0$ which means that $\hat{x}_n = x_n$. It trivially seen from Lemma 2 assuming that $x(t_i) \neq 0$ that

$$x_{n-1} = \hat{x}_{n-1} + x_{n-1}^d = \hat{x}_{n-1} + f^{2n-4}(-\lambda_n) \frac{1}{k} \hat{x}_n \quad (61)$$

which implies that

$$\hat{g}_{n-1,n} = g_{n-1,n} = -\lambda_n. \quad (62)$$

We can then calculate from the definition (33) of x_{n-2}^d and (60)–(62)

$$\begin{aligned} x_{n-2} &= \hat{x}_{n-2} + x_{n-2}^d \\ &= \hat{x}_{n-2} + f^{2n-6} \left(g_{n-2,n-1} \frac{x_{n-1}}{k} + g_{n-2,n} \frac{x_n}{k^2} \right) \\ &= \hat{x}_{n-2} + f^{2n-6} \left[g_{n-2,n-1} \frac{1}{k} \hat{x}_{n-1} \right. \\ &\quad \left. + (g_{n-2,n} + g_{n-2,n-1} \hat{g}_{n-2,n} f^{2n-4}) \frac{x_n}{k^2} \right] \\ &= \hat{x}_{n-2} + f^{2n-6} \left[\hat{g}_{n-2,n-1} \frac{1}{k} \hat{x}_{n-1} + \hat{g}_{n-2,n} \frac{1}{k^2} \hat{x}_n \right] \end{aligned} \quad (63)$$

where

$$\hat{g}_{n-2,n-1} = g_{n-2,n-1} \quad (64)$$

$$\hat{g}_{n-2,n} = g_{n-2,n} + g_{n-2,n-1} g_{n-1,n} f^{2n-4}. \quad (65)$$

The rest of the proof will be given by induction. Assume that (47) is satisfied for $j+1, j+2, \dots, n-1$, i.e.,

$$x_{m+1} = \hat{x}_{m+1} + f^{2m} \sum_{r=m+2}^n \hat{g}_{m+1,r} \frac{1}{k^{r-m-1}} \hat{x}_r \quad (66)$$

where $m \in \{j, j+1, \dots, n-2\}$. We then find from the definition (33) of x_j^d assuming that $x(t_i) \neq 0$

$$\begin{aligned} x_j &= \hat{x}_j + x_j^d \\ &= \hat{x}_j + f^{2j-2}(t) \sum_{r=j+1}^n g_{jr} \frac{1}{k^{r-j}} x_r \\ &= \hat{x}_j + f^{2j-2}(t) \sum_{r=j+1}^n g_{jr} \frac{1}{k^{r-j}} \\ &\quad \times \left[\hat{x}_r + f^{2r-2} \sum_{m=r+1}^n \hat{g}_{rm} \frac{1}{k^{m-r}} \hat{x}_m \right] \\ &= \hat{x}_j + f^{2j-2}(t) \sum_{r=j+1}^n g_{jr} \\ &\quad \times \left[1 + f^{2r-2} \sum_{m=j+1}^{r-1} \hat{g}_{mr} \right] \frac{1}{k^{r-j}} \hat{x}_r \\ &= \hat{x}_j + f^{2j-2}(t) \sum_{r=j+1}^n \hat{g}_{jr} \frac{1}{k^{r-j}} \hat{x}_r \end{aligned} \quad (67)$$

where

$$\hat{g}_{jr} = g_{jr} \left[1 + f^{2r-2} \sum_{m=j+1}^{r-1} \hat{g}_{mr} \right]. \quad (68)$$

Equation (67) shows that if (47) is satisfied for $j+1, \dots, n$, then (47) is satisfied for j as well. Thus, (61)–(62) and (67) imply that (47) is satisfied for all $\{2, \dots, n-1\}$. We see that \hat{g}_{jr} is given by g_{mr} and q_i which are known. Because of the smoothness and boundedness of $f(t), g_{jr}(t)$ and $\hat{g}_{mr}(t)$ for $m \in \{j+1, \dots, n-1\}$, \hat{g}_{jr} will be smooth and bounded as well.

Proof of Theorem 1: The point of time t_i denotes the largest term in the sequence (t_0, t_2, \dots) such that $t \geq t_i$. If $x(t_i) = 0$ then $[u_1(t), u_2(t)]^T = 0$ from the control law (22) and (31) for $t \in [t_i, t_{i+1})$. From the model (15) we then see that $x(t) = 0$, for $t \in [t_i, t_{i+1})$, implying that $[u_1(t), u_2(t)]^T = 0$ for $t \in [t_{i+1}, t_{i+2})$, etc. and $x(t) = 0$ for all $t \geq t_i$. In that case, relation (48) is satisfied for all $t \geq t_i$ for any function $h_z(\cdot)$ of class \mathcal{K} . In the following, we analyze the case where $x(t_i) \neq 0$. The norm $\|\cdot\|$ denotes the one-norm (18).

From Lemma 4 we have that for $j \in \{2, \dots, n-1\}$

$$\begin{aligned} x_j(t) &= \hat{x}_j(x(t), t) + f^{2j-2}(t) \\ &\quad \times \sum_{r=j+1}^n \hat{g}_{jr}(t) \frac{1}{k^{r-j}(x(t), t)} \hat{x}_r(x(t), t) \end{aligned}$$

where the functions $\hat{g}_{jr}(t)$ are bounded, say by the constant G , such that $|\hat{g}_{jr}(t)| \leq G$ for all $t \geq t_0$. Since $|f(t)| \leq 1$, Property P2) of $f(t)$, we get

$$\begin{aligned} |x_j(t)| &\leq |\hat{x}_j(x(t), t)| + \sum_{r=j+1}^n G \frac{1}{|k(x(t), t)|^{r-j}} |\hat{x}_r(x(t), t)|, \\ j &\in \{2, \dots, n-1\}. \end{aligned} \quad (69)$$

Now, we will show that there are constants κ_j such that

$$|k(x(t_i))| = K \text{ or } |k(x(t_i))| \geq \kappa_j |x_j(x(t), t)|^{\frac{1}{2^{j-1}}} \quad (70)$$

for all $j \in \{3, \dots, n\}$. From (23) we find that if $|k(x(t_i))| < K$ then

$$|k(x(t_i))| = \|x_1(t_i)\| + G(\|z(t_i)\|)\beta.$$

Combining this with the definition of $G(\cdot)$, (24), implies

$$\begin{aligned} |k(x(t_i))| &\geq G(\|z(t_i)\|)\beta = \beta \kappa \|z(t_i)\|^{\frac{1}{2^{n-4}}} \\ &\geq \beta \kappa \|\hat{z}_j(t_i)\|^{\frac{1}{2^{n-4}}} \\ j &\in \{2, \dots, n\}. \end{aligned} \quad (71)$$

Lemma 3 implies

$$\begin{aligned} |k(x(t_i))| &\geq \beta \kappa \left(\frac{1}{c_j} |\hat{x}_j(x(t), t)| \right)^{\frac{1}{2^{j-1}}} \\ &= \beta \kappa c_j^{\frac{1}{2^{j-1}}} |\hat{x}_j(x(t), t)|^{\frac{1}{2^{j-1}}}, \quad j \in \{3, \dots, n\}. \end{aligned} \quad (72)$$

Therefore, (70) is satisfied by defining $\kappa_j = \beta \kappa c_j^{\frac{1}{2^{j-1}}}$.

From (70) and (69) we then get

$$\begin{aligned} |x_j(t)| &\leq |\hat{x}_j(x(t), t)| + G \sum_{r=j+1}^n \left[\frac{1}{K^{r-j}} |\hat{x}_r(x(t), t)| \right. \\ &\quad \left. + \frac{1}{\kappa_r^{r-j}} |\hat{x}_r(x(t), t)|^{1-\frac{r-j}{2^{n-4}}} \right] \end{aligned} \quad (73)$$

for all $j \in \{2, \dots, n-1\}$. From Lemma 3 we get, with $t_p = t_0$

$$\begin{aligned} |x_j(t)| &\leq c_j \|z_j(t_0)\| e^{-\gamma_j(t-t_0)} \\ &\quad + G \sum_{r=j+1}^n \left[\frac{1}{K^{r-j}} c_r \|z_r(t_0)\| e^{-\gamma_r(t-t_0)} \right. \\ &\quad \left. + \left(\frac{c_r}{K^{r-j}} \|z_r(t_0)\| e^{-\gamma_r(t-t_0)} \right)^{1-\frac{r-j}{2n-4}} \right] \\ &\leq h_j(\|z(t_0)\|) e^{-\zeta_j(t-t_0)} \end{aligned}$$

for all $j \in \{2, \dots, n-1\}$, since $\|z(t_0)\| \geq \|z_r(t_0)\|$, where

$$\begin{aligned} h_j(q) &= c_j q + G \sum_{r=j+1}^n \left[\frac{1}{K^{r-j}} c_r q + \left(\frac{c_r}{K^{r-j}} q \right)^{1-\frac{r-j}{2n-4}} \right] \\ \zeta_j &= \min_r \left\{ \gamma_r \left(1 - \frac{r-j}{2n-4} \right) \right\}, \quad r \in \{j, \dots, n\}. \end{aligned} \quad (74)$$

From (44)–(46) we see that

$$0 < \gamma_n < \gamma_{n-1} < \dots < \gamma_2 < \lambda_2.$$

Equation (74) then implies that

$$\zeta_j = \gamma_n \left(1 - \frac{n-j}{2n-4} \right).$$

Note that the function $h_j(q)$ is of class \mathcal{K} . Since $x_n^d \triangleq 0$ we have from Lemma 3

$$|x_n(t)| = |\tilde{x}_n(x(t), t)| \leq c_n \|z(t_0)\| e^{-\gamma_n(t-t_0)}.$$

We define $h_n(\|z(t_0)\|) = c_n \|z(t_0)\|$ and $\zeta_n = \gamma_n$ such that from Lemma 3

$$|x_n(t)| = |\tilde{x}_n(t)| \leq h_n(\|z(t_0)\|) e^{-\gamma_n(t-t_0)}.$$

The proof is completed by noting that

$$\begin{aligned} \|z(t)\| &= |x_2(t)| + \dots + |x_n(t)| \\ &\leq \sum_{j=2}^n h_j(\|z(t_0)\|) e^{-\zeta_j(t-t_0)} = h_z(\|z(t_0)\|) e^{-\gamma_z(t-t_0)} \end{aligned}$$

where

$$h_z(q) = \sum_{j=2}^n h_j(q) \quad (75)$$

$$\gamma_z = \min\{\zeta_2, \dots, \zeta_n\} = \zeta_2 = \frac{\gamma_n}{2}. \quad (76)$$

□

Proof of Lemma 5: The outline of the proof is to show that we can find a small enough neighborhood Ω of the origin such that the controller u_1 does not get saturated due to the convergence of $z(t)$ from Theorem 1. It will then be shown that the control law $u_1 = k(x(t_i))f(t)$ is chosen such that $x_1(t)$ converges to zero as $z(t)$ converges. A function $h(\cdot; T)$ will be constructed. The norm $\|\cdot\|$ denotes the one-norm (18).

Let the neighborhood be given by

$$\Omega = \left\{ x \mid |x_1| < \frac{K}{2\beta}, G(h_z(\|z\|)) < \frac{K}{2\beta} \right\}$$

where $h_z(\cdot)$ is a function of class \mathcal{K} from Theorem 1 and $z = [x_2, \dots, x_n]^T$ (16). The function $G(\cdot)$ is defined in (24),

and β is defined in (25). By induction, we will show that if $x(t_0) \in \Omega$, then

$$|k(x(t_i))| < K, \quad \forall t_i \in \{t_0, t_1, \dots\} \quad (77)$$

which implies from the definition of $k(\cdot)$, (23)

$$\begin{aligned} k(x(t_i)) &= -[x_1(t_i) + \text{sgn}(x_1(t_i))G(\|z(t_i)\|)]\beta, \\ \forall t_i &\in \{t_0, t_1, \dots\}. \end{aligned} \quad (78)$$

Note from Theorem 1 that

$$\|z(t)\| \leq h_z(\|z(t_0)\|), \quad \forall t \geq t_0.$$

Since $G(\cdot)$ is of class \mathcal{K} , (24), this implies

$$G(\|z(t)\|) \leq G(h_z(\|z(t_0)\|)), \quad \forall t \geq t_0.$$

Assume for a $t_m \in \{t_0, t_1, \dots\}$ that

$$k(x(t_m)) = -[x_1(t_m) + \text{sgn}(x_1(t_m))G(\|z(t_m)\|)]\beta.$$

Integrating $\dot{x}_1 = u_1 = k(x(t_m))f(t)$ from t_m to t_{m+1} then gives

$$x_1(t_{m+1}) = -\text{sgn}(x_1(t_m))G(\|z(t_m)\|) \quad (79)$$

which implies that

$$|x_1(t_{m+1})| = G(\|z(t_m)\|) \leq G(h_z(\|z(t_0)\|)).$$

By assumption, $x(t_0) \in \Omega$ which implies that $G(h_z(\|z(t_0)\|)) < \frac{K}{2\beta}$ and therefore

$$|x_1(t_{m+1})| < \frac{K}{2\beta}.$$

Since

$$[|x_1(t_{m+1})| + G(\|z(t_{m+1})\|)]\beta < \left(\frac{K}{2\beta} + \frac{K}{2\beta} \right)\beta = K$$

then from the definition of $k(\cdot)$, (23)

$$k(x(t_{m+1})) = -[x_1(t_{m+1}) + \text{sgn}(x_1(t_{m+1}))G(\|z(t_{m+1})\|)]\beta.$$

Equations (77) and (78) are proved by noting that since $x(t_0) \in \Omega$, then

$$[|x_1(t_0)| + G(\|z(t_0)\|)]\beta < \left(\frac{K}{2\beta} + \frac{K}{2\beta} \right)\beta = K$$

which implies that

$$k(x(t_0)) = -[x_1(t_0) + \text{sgn}(x_1(t_0))G(\|z(t_0)\|)]\beta.$$

Integrating $\dot{x}_1 = u_1$ in (15) from t_i to $t < t_{i+1}$ with $u_1 = k(x(t_i))f(t)$ and k constant gives

$$\begin{aligned} |x_1(t)| &\leq |x_1(t_i)| + |k| \int_{t_i}^t f(\tau) d\tau \\ &\leq |x_1(t_i)| + |k| \int_{t_i}^{t_{i+1}} f(\tau) d\tau \\ &= |x_1(t_i)| + |x_1(t_{i+1}) - x_1(t_i)| \\ &\leq 2|x_1(t_i)| + |x_1(t_{i+1})|. \end{aligned} \quad (80)$$

once $x(t_0) \in \Omega$ then $k(x(t_i))$ is given by (78) and we get from (79) and (80)

$$r_1(t) \leq 2G(\|z(t_{i-1})\|) + G(\|z(t_i)\|), \quad i \in \{1, 2, \dots\}. \quad (81)$$

From Theorem 1 we have that

$$\|z(t_k)\| \leq h_z(\|z(t_0)\|)e^{-\gamma_z(t_k-t_0)}. \quad (82)$$

From the definition of $G(\cdot)$, (24), we have

$$G(ae^{-bt}) = \kappa(ae^{-bt})^{\frac{1}{2n-4}} = e^{\frac{-bt}{2n-4}} G(a) \quad (83)$$

where a and b are positive constants. We denote for simplicity

$$q = h_z(\|z(t_0)\|).$$

Equation (83) combined with (81) and (82) implies

$$\begin{aligned} |x_1(t)| &\leq 2G(qe^{-\gamma_z(t_{i-1}-t_0)}) + G(qe^{-\gamma_z(t_i-t_0)}) \\ &= 2G(q)e^{\frac{-\gamma_z}{2n-4}(t_{i-1}-t_0)} + G(q)e^{\frac{-\gamma_z}{2n-4}(t_i-t_0)} \\ &\leq 3G(q)e^{\frac{-\gamma_z}{2n-4}(t_{i-1}-t_0)}. \end{aligned}$$

By convention we have for all $i \in \{1, 2, \dots\}$

$$t_{i-1} = t_{i+1} - 2T \geq t - 2T$$

since $t \in [t_i, t_{i+1})$. This implies

$$|x_1(t)| \leq 3G(q)e^{\frac{-\gamma_z 2T}{2n-4}} e^{\frac{-\gamma_z(t-t_0)}{2n-4}}, \quad t \geq t_1. \quad (84)$$

If $t \in [t_0, t_1)$ we find the following bound on $|x_1(t)|$ by integrating $\dot{x}_1 = k(x(t_0))f(t)$ where k is given by (78) and $f(t) \geq 0$

$$\begin{aligned} |x_1(t)| &\leq |x_1(t_0)| + \int_{t_0}^t f(\tau) d\tau \\ &\leq 2|x_1(t_0)| + G(\|z(t_0)\|). \end{aligned}$$

Since $t \in [t_0, t_1)$ this implies

$$|x_1(t)| \leq [2|x_1(t_0)| + G(\|z(t_0)\|)]e^{-\frac{\gamma_z(t-t_0)}{2n-4}}, \quad t \in [t_0, t_1). \quad (85)$$

Combining (84) and (85) and using $t_1 - t_0 = T$ implies

$$\begin{aligned} |x_1(t)| &\leq \{3G(h_z(\|z(t_0)\|))e^{\frac{\gamma_z 2T}{2n-4}} \\ &\quad + [2|x_1(t_0)| + G(\|z(t_0)\|)]e^{\frac{\gamma_z T}{2n-4}}\}e^{\frac{-\gamma_z(t-t_0)}{2n-4}} \\ &\leq h_1(\|x(t_0)\|; T)e^{\frac{-\gamma_z(t-t_0)}{2n-4}}, \quad t \geq t_0 \end{aligned} \quad (86)$$

where

$$\begin{aligned} h(\|x(t_0)\|; T) &= 3G(h_z(\|x(t_0)\|))e^{\frac{\gamma_z 2T}{2n-4}} \\ &\quad + [2\|x(t_0)\| + G(\|x(t_0)\|)]e^{\frac{\gamma_z T}{2n-4}}. \end{aligned}$$

Since $G(\cdot)$ and $h_z(\cdot)$ are functions of class \mathcal{K} , $h_1(\cdot; T)$ is also of class \mathcal{K} , (1). The proof is completed by noting from Theorem 1 and (86) that

$$\begin{aligned} \|x(t)\| &= |x_1(t)| + \|z(t)\| \\ &\leq h_1(\|x(t_0)\|; T)e^{\frac{-\gamma_z(t-t_0)}{2n-4}} + h_z(\|z(t_0)\|)e^{-\gamma_z(t-t_0)} \\ &\leq [h_1(\|x(t_0)\|; T) + h_z(\|z(t_0)\|)]e^{\frac{-\gamma_z(t-t_0)}{2n-4}} \\ &= h(\|x(t_0)\|; T)e^{-\gamma(t-t_0)}, \quad t \geq t_0 \end{aligned} \quad (87)$$

where

$$\begin{aligned} h(\|x(t_0)\|; T) &= h_1(\|x(t_0)\|) + h_z(\|z(t_0)\|), \\ \gamma &= \frac{\gamma_z}{2n-4} \end{aligned}$$

where γ_z is given from Theorem 1.

Proof of Theorem 2. The theorem will be proved by showing that $z(t) = [x_2(t), \dots, x_n(t)]^T$ and $x_1(t)$ are bounded and reach the neighborhood Ω defined in Lemma 5 in finite time. After this finite time, exponential convergence of $x(t)$ is ensured by Lemma 5. A function $h_z(\|z(t_0)\|; T)$ is constructed by using expressions for the finite time and the bounds on $\|x(t)\|$ and $h(\|x(t_0)\|; T)$ in Lemma 5. The norm $\|\cdot\|$ denotes the one-norm (18).

We will seek a time $\tau(\|x(t_0)\|)$ such that for all $t \geq \tau$, $x(t) \in \Omega$. First, define the time function $T_z: R_+ \rightarrow R_+$ as

$$T_z(q) \triangleq \begin{cases} \frac{1}{\gamma_z} \ln G(h_z(h_z(q))) \frac{2\beta}{K}, & G(h_z(h_z(q))) \geq \frac{K}{2\beta} \\ 0, & G(h_z(h_z(q))) < \frac{K}{2\beta} \end{cases} \quad (89)$$

Consider $G(h_z(\|z(t)\|))$. By inserting (48) and accounting for $\|z\| \leq \|x\|$ and the fact that $G(\cdot)$ and $h_z(\cdot)$ are of class \mathcal{K} , it is seen that

$$\forall t > T_z(\|x(t_0)\|) + t_0, \quad G(h_z(\|z(t)\|)) < \frac{K}{2\beta}. \quad (90)$$

Then, integrating $\dot{x}_1 = u_1 = k(x(t))f(t)$ from t_i to t_{i+1} gives

$$x_1(t_{i+1}) = \begin{cases} -\text{sgn}(x_1(t_i))G(\|z(t_i)\|), & |k(x(t_i))| < K \\ x_1(t_i) - \text{sgn}(x_1(t_i))\frac{K}{\beta}, & |k(x(t_i))| = K \end{cases} \quad (91)$$

Define the time function $\tau: R_+ \rightarrow R_+$ as

$$\tau(q) = \max \left\{ T_z(q), \frac{\beta}{K} q \right\}. \quad (92)$$

One can show from (90) and (91) that for all $t_m \in \{t_0, t_1, \dots\}$ which satisfies $t_m > t_0 + \tau(\|x(t_0)\|)$

$$|x_1(t_m)| < \frac{K}{2\beta}. \quad (93)$$

By integrating $\dot{x}_1 = k(x(t_m))f(t)$ one can show that $x_1(t)$ is a linear interpolation between $x_1(t_m)$ and $x_1(t_{m+1})$ for all $t \in [t_m, t_{m+1})$. From the definition of Ω , (50), and from (90) and (93) we find

$$\forall t > t_0 + \tau(\|x(t_0)\|), \quad x(t) \in \Omega.$$

Denote by t_p the smallest element in the sequence $\{t_0, t_1, \dots\}$ such that $t_p > t_0 + \tau(\|x(t_0)\|)$. From Lemma 5 it then follows that

$$\forall t \geq t_p, \quad \|x(t)\| \leq h(\|x(t_p)\|; T)e^{-\gamma(t-t_p)}. \quad (94)$$

To include $t \in [t_0, t_p)$, an expression for the maximum of $\|x(t)\|$ must be found.

We can show from (91) by using the convergence property of $z(t)$, Theorem 1, that

$$\begin{aligned} \max_{t \geq t_0} \|x_1(t)\| &\leq \max\{\|x_1(t_0)\|, G(h_z(\|z(t_0)\|))\} \\ &\leq \max\{\|x(t_0)\|, G(h_z(\|x(t_0)\|))\} \\ &\triangleq x_{1m}(\|x(t_0)\|) \end{aligned} \quad (95)$$

since $\|x(t_0)\| \geq \|z(t_0)\|$, and $h_z(\cdot)$ are of class \mathcal{K} . Here, we have defined the function $x_{1m}: R_+ \rightarrow R_+$ which is of class \mathcal{K} .

From (95) and Lemma 5 we have

$$\begin{aligned} \|x(t)\| &\leq x_{1m}(\|x(t_0)\|) + h_z(\|x(t_0)\|) \\ &\triangleq x_{mx}(\|x(t_0)\|), \quad \forall t \geq t_0. \end{aligned}$$

Here, we have defined the function $x_{mx}: R_+ \rightarrow R_+$ which is of class \mathcal{K} . The bound in (94) can then be extended to an exponential bound for all $t \geq t_0$ and all $x(t_0) \in R^n$

$$\|x(t)\| \leq [h(x_{mx}(q); T) + x_{mx}(q)]e^{\gamma\tau(q)}e^{-\gamma(t-t_0)}$$

where q denotes $\|x(t_0)\|$. From the definition of $\tau(\cdot)$, (92), we then obtain for all $t \geq t_0$

$$\forall x(t_0) \in R^n, \quad \|x(t)\| \leq h_x(\|x(t_0)\|; T)e^{-\gamma(t-t_0)}$$

where the class \mathcal{K} function $h_x: R_+ \rightarrow R_+$ is defined as

$$h_x(q; T) \triangleq [h(x_{mx}(q); T) + x_{mx}(q)]e^{\gamma\tau(q)}$$

and q denotes $\|x(t_0)\|$. \square

ACKNOWLEDGMENT

The authors would like to thank Dr. C. Canudas de Wit for valuable discussions on the stabilization of nonholonomic systems and Dr. C. Samson for comments on the stabilization about any configuration.

REFERENCES

- [1] J. Barraquand and J.-C. Latombe, "On nonholonomic mobile robots and optimal maneuvering," *Revue d'Intelligence Artificielle*, vol. 3, no. 2, pp. 77-103, 1989.
- [2] A. M. Bloch and N. H. McClamroch, "Control of mechanical systems with classical nonholonomic constraints," in *Proc. 28th Conf. Decis. Contr.*, Tampa, FL, Dec. 1989, pp. 201-205.
- [3] A. M. Bloch, N. H. McClamroch, and M. Reyhanoglu, "Controllability and stabilizability properties of a nonholonomic control system," in *Proc. 29th Conf. Decis. Contr.*, Honolulu, HI, Dec. 1990, pp. 1312-1314.
- [4] R. W. Brockett, "Asymptotic stability and feedback stabilization," in *Differential Geometric Control Theory*, R. W. Brockett, R. S. Millman, and H. J. Sussman, Eds. Boston: Birkhauser, 1983, pp. 181-208.
- [5] C. Canudas de Wit and O. J. Sordalen, "Exponential stabilization of mobile robots with nonholonomic constraints," *IEEE Trans. Automat. Contr.*, vol. 37, no. 11, pp. 1791-1797, 1992.
- [6] J.-M. Coron, "Global asymptotic stabilization for controllable systems without drift," *Math. Contr., Sig., Syst.*, vol. 5, pp. 295-315, 1991.
- [7] J.-M. Coron and J.-P. Pomet, "A remark on the design of time-varying stabilizing feedback laws for controllable systems without drift," in *Proc. Nonlinear Contr. Syst. Design Symp.*, Bordeaux, France, June 1992, pp. 413-417.
- [8] B. d'Andrea-Novell, G. Bastin, and G. Campion, "Modeling and control of nonholonomic wheeled mobile robots," in *Proc. 1991 IEEE Int. Conf. Robotics Automation*, Sacramento, CA, Apr. 1991, pp. 1130-1135.
- [9] L. Gurvits, "Averaging approach to nonholonomic motion planning," in *Proc. 1992 IEEE Int. Conf. Robotics Automation*, Nice, France, May 1992, pp. 2541-2546.
- [10] L. Gurvits and Z. X. Li, "Smooth time-periodic feedback solutions for nonholonomic motion planning," in *Progress in Nonholonomic Motion Planning*, Z. X. Li and J. Canny, Eds. Norwell, MA: Kluwer, 1993, pp. 53-108.
- [11] W. Hahn, *Stability of Motion*. New York: Springer-Verlag, 1967.
- [12] G. Jacob, "Motion planning by piecewise constant or polynomial inputs," in *Proc. nonlinear contr. syst. design symp.*, Bordeaux, France, June 1992, pp. 628-633.
- [13] M. Kawski, "Homogeneous stabilizing feedback laws," *Contr. Theory Advanced Tech.*, vol. 6, no. 4, pp. 497-516, 1990.
- [14] H. K. Khalil, *Nonlinear Systems*. New York: Macmillan, 1992.
- [15] G. Lafferriere, "A general strategy for computing controls of systems without drift," in *Proc. 30th Conf. Decis. Contr.*, Brighton, England, Dec. 1991, pp. 1115-1120.
- [16] G. Lafferriere and H. J. Sussmann, "Motion planning for controllable systems without drift," in *Proc. 1991 IEEE Int. Conf. Robotics and Automation*, Sacramento, California, Apr. 1991, pp. 1148-1153.
- [17] J.-P. Laumond, "Feasible trajectories for mobile robots with kinematic and environment constraints," in *Proc. Int. Conf. Intelligent Autonomous Syst.*, Amsterdam, The Netherlands, 1986, pp. 346-354.
- [18] ———, "Finding collision-free smooth trajectories for a nonholonomic mobile robot," in *10th Int. Joint Conf. Artificial Intelligence*, Milano, Italy, 1987, pp. 1120-1123.
- [19] J.-P. Laumond, M. Taix, and P. Jacobs, "A motion planner for car-like robots based on a mixed global/local approach," in *Proc. Int. Conf. Intelligent Robots Syst.*, Japan, 1990, pp. 765-773.
- [20] R. T. M'Closkey and R. M. Murray, "Convergence rates for nonholonomic systems in power form," in *Proc. Amer. Contr. Conf.*, San Francisco, CA, June 1993, pp. 2967-2972.
- [21] R. K. Miller and A. N. Michel, *Ordinary Differential Equations*. New York: Academic, 1982.
- [22] S. Monaco and D. Normand-Cyrot, "An introduction to motion planning under multirate control," in *Proc. 31st Conf. Decis. Contr.*, Tucson, AZ, Dec. 1992, pp. 1780-1785.
- [23] R. M. Murray, "Nilpotent bases for a class of nonintegrable distributions with applications to trajectory generation for nonholonomic systems," California Inst. Tech., Pasadena, Tech. Memo, CIT-CDS 92-002, Oct. 1992.
- [24] R. M. Murray and S. S. Sastry, "Steering nonholonomic systems using sinusoids," in *Proc. 29th Conf. Decis. Contr.*, Honolulu, HI, Dec. 1990, pp. 2097-2101.
- [25] ———, "Steering nonholonomic systems in chained form," in *Proc. 30th Conf. Decis. Contr.*, Brighton, England, Dec. 1991, pp. 1121-1126.
- [26] R. M. Murray and S. S. Sastry, "Nonholonomic motion planning: Steering using sinusoids," *IEEE Trans. Automat. Contr.*, vol. 38, no. 5, pp. 700-716, 1993.
- [27] R. M. Murray, G. Walsh, and S. S. Sastry, "Stabilization and tracking for nonholonomic control systems using time-varying state feedback," in *Proc. Nonlinear Contr. Syst. Design Symp.*, Bordeaux, France, June 1992, pp. 182-187.
- [28] J.-P. Pomet, "Explicit design of time-varying stabilizing control laws for a class of controllable systems without drift," *Syst. Contr. Lett.*, vol. 18, no. 2, pp. 147-158, 1992.
- [29] C. Samson, "Time-varying feedback stabilization of nonholonomic car-like mobile robots," INRIA-Sophia Antipolis, Tech. Rep. 1515, Sep. 1991.
- [30] ———, "Velocity and torque feedback control of a nonholonomic cart," in *Advanced Robot Contr. Proc. Int. Workshop Nonlinear Adaptive Contr: Issues in Robotics*, France, Nov. 1990, pp. 125-151.
- [31] C. Samson and K. Ait-Adberrahim, "Feedback stabilization of a nonholonomic wheeled mobile robot," in *Proc. Int. Conf. Intelligent Robots Syst.*, Japan, 1990.
- [32] ———, "Mobile robot control, part 1: Feedback control of nonholonomic wheeled cart in cartesian space," INRIA-Sophia Antipolis, Tech. Rep. 1288, Oct. 1990.
- [33] O. J. Sordalen, "Conversion of the kinematics of a car with n trailers into a chained form," in *Proc. 1993 IEEE Int. Conf. Robotics and Automation*, Atlanta, GA, May 1993, pp. 382-387.
- [34] A. R. Teel, R. M. Murray, and G. Walsh, "Nonholonomic control systems: From steering to stabilization with sinusoids," in *Proc. 31st Conf. Decis. Contr.*, Tucson, AZ, Dec. 1992, pp. 1603-1609.
- [35] D. Tilbury, R. Murray, and S. Sastry, "Trajectory generation for the n -trailer problem using goursat normal form," University of California, Berkeley, Tech. Rep. UCB/ERL M93/12, Feb. 1993.



Ole Jakob Sørдалen (S'92-M'93) was born in Kragerø, Norway, in 1965. He received the Diploma Engineer (Siv. Ing.) degree in 1988 and the Dr. Ing. degree in 1993 in electrical engineering and Computer Sciences at the Norwegian Institute of Technology, Trondheim, Norway.

Dr. Sørдалen is currently a Researcher at ABB Corporate Research, Norway. He has held visiting appointments at Laboratoire d'Automatique de Grenoble, France, University of California, Berkeley, and University of Tokyo. His research interests

include nonlinear control of mechanical systems and nonholonomic systems.



Olav Egeland (S'85-M'86) was born in Leirvik, Norway, in 1959. He received the Siv. Ing. degree in 1984 and the Dr. Ing. degree in 1987 in Electrical Engineering at the Norwegian Institute of Technology, Trondheim, Norway.

Dr. Egeland became Assistant Professor in 1987 and Professor of Robotics in 1989 at the Department of Engineering Cybernetics, the Norwegian Institute of Technology. In the academic year 1988/89 he was a Visiting Scientist at the German Aerospace Research Establishment (DLR) in Oberpfaffenhofen, Germany.

His research interests include control of manipulators, surface vessels, flexible structures and nonholonomic vehicles.

Global Total Least Squares Modeling of Multivariable Time Series

Berend Roorda and Christiaan Heij, *Member, IEEE*

Abstract—In this paper we present a novel approach for the modeling of multivariable time series. The model class consists of linear systems, i.e., the solution sets of linear difference equations. Restricting the model order, the aim is to determine a model with minimal l_2 -distance from the observed time series. Necessary conditions for optimality are described in terms of state-space representations. These conditions motivate a relatively simple iterative algorithm for the nonlinear problem of identifying optimal models. Attractive aspects of the proposed method are that the model error is measured globally, it can be applied for multi-input, multi-output systems, and no prior distinction between inputs and outputs is required. We give an illustration by means of some numerical simulations.

I. INTRODUCTION

THE basic problem in time series modeling is to find a reasonably simple model which gives a sufficiently accurate description of the data. Procedures which have been developed for this problem differ in the specification of the model class and in the way the complexity and accuracy of models is evaluated.

In this paper we will restrict our attention to models described by linear difference equations with fixed coefficients and finite lags. Within this classical setting several modeling procedures have been developed. For an overview, we refer to the textbooks [1], [4], [6]. Well-known examples are the least squares identification of input-output systems in polynomial form and, more generally, the maximum likelihood identification of ARMAX systems, i.e., input-output models where the disturbances follow a moving average process. These and most other methods require that several structural aspects of the model should be specified *a priori*.

- 1) The number of equations, that is, the number of inputs and outputs of the system, and the decomposition of the system variables into inputs and outputs.
- 2) The orders of each of the equations, that is, the so-called structural indices of the system.
- 3) The stochastic properties of the disturbances, in particular the joint correlation structure between inputs, outputs, and disturbances.
- 4) The choice of a canonical parameterization, to avoid problems of nonidentifiable parameters.

Manuscript received July 22, 1993; revised February 5, 1994 and May 25, 1994. Recommended by Associate Editor, S. P. Meyn. The work was supported in part by Grant SC1*-CT92-0779 on System Identification of the Science Program of the Commission of the European Community.

B. Roorda is with the Tinbergen Institute, Erasmus University Rotterdam, Oostmaaslaan 950-952, 3063 DM Rotterdam, Netherlands.

C. Heij is with the Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, Netherlands.

IEEE Log Number 9406098.

The identification procedure that we propose in this paper differs in some crucial aspects from the methods just described. Our aim is to decompose a given multivariable time series, denoted by w , into two parts, i.e.,

$$w = \hat{w} + \tilde{w} \quad (1)$$

where \hat{w} represents a regular part and \tilde{w} the corresponding deviation. The aim is to keep the approximation error as small as possible, under the condition that the approximating time series \hat{w} is sufficiently regular. To make this more explicit we next describe our notions of regularity and model error.

A time series is called regular if it satisfies linear, time-invariant difference equations of finite lag. Let q denote the number of system variables, p the number of independent equations, and n the total lag, i.e., the sum of the lags of the individual equations. Time series are more regular when $m := q - p$ and n are smaller, i.e., the more equations they satisfy and the smaller the number of initial conditions. This can also be formulated as follows. Define the complexity of a linear, time-invariant, finite dimensional system by the pair (m, n) , where m is the number of system inputs and n the (minimal) number of state variables. A system is called less complex if it has fewer inputs, i.e., unexplained variables, and if it has less states, i.e., initial degrees of freedom. Then a time series is more regular if it can be generated by a less complex system.

The model error is evaluated as follows. For expository reasons we restrict ourselves in this paper to time series which are specified over the infinite time axis \mathbb{Z} and which are square summable, i.e., we assume that $w \in l_2^q$. The main results in this paper can be extended for modeling time series observed over a finite time interval, but we will not treat this issue here to simplify the presentation. The error in approximating an observed time series w by a regular part \hat{w} is measured by the l_2 -norm of the deviation $\tilde{w} = w - \hat{w}$, denoted by $\|\tilde{w}\| := \{\sum_{t=-\infty}^{\infty} \tilde{w}(t)^T \tilde{w}(t)\}^{1/2}$.

In our approach we will assume that the required regularity of the approximating time series \hat{w} has been specified *a priori*. Stated otherwise, we impose an upper bound on the complexity of the system that can generate the approximation. We denote by $\mathcal{B}^{q,m,n}$ the set of all time series that can be generated by systems with m inputs, $q - m$ outputs, and n states, i.e., all time series that satisfy $q - m$ independent linear, time-invariant difference equations with total lag n . Under this restriction we wish to minimize the approximation error as defined above, i.e.,

$$\min\{\|w - \hat{w}\|; \hat{w} \in \mathcal{B}^{q,m,n}\}. \quad (2)$$

Solving this problem for different values of (m, n) gives an impression of the involved trade-off between the required regularity and the resulting approximation error. The search for an acceptable model complexity is facilitated by the fact that the error decreases for increasing (m, n) , as $\mathcal{B}^{q \times m \times n} \subset \mathcal{B}^{(m_2 \times n_2) \times m_1 \times n_1}$ if $m_1 \leq m_2$ and $n_1 \leq n_2$. Note that this criterion follows for deviations in all the system variables and that the error is not measured locally (e.g., as a prediction error) but as the global l_2 -distance between the observation and the regular part. Therefore we give the name 'global total least squares' to the identification criterion (2).

As compared to classical procedures, our formulation of the time series modeling problem involves less *a priori* specifications. With respect to the four structural aspects discussed before, our approach has the following features:

- 1) The number of inputs and outputs is specified *a priori* but the system variables are all treated alike so that there is no need for a specification of inputs and outputs.
- 2) The structural indexes need not be specified but only the total lag and this can be varied easily.
- 3) The problem formulation involves no stochastic specifications although these may be incorporated by adjusting the norm on l_2 .
- 4) The criterion is nonparametric so that any representation may be chosen as it suits.

This paper has the following structure. To give some feeling for the global total least squares problem we first describe the well-known and relatively simple case of static total least squares in Section II. The basis for our modeling theory is the behavioral approach to systems and this is briefly discussed in Section III. For further treatment of the behavioral approach in systems theory we refer to [10] and [11]. In Sections IV and V we develop a highly structured type of system representations: isometric state representations which form the cornerstone of our modeling theory. Section VI concerns the question of how to determine an optimal approximation of an observed time series within a given model. This linear optimization problem is solved by a projection algorithm which was introduced in [12]. In Section VII we treat the problem of determining an optimal system. This is a nonlinear optimization problem over a nonconvex set. We propose three model improvement constructions based on the results in Section VI. These constructions are used in Section VIII to estimate locally optimal models. In Section IX we describe three simulation experiments that illustrate the use of the global total least squares method, and Section X contains some conclusions.

II. STATIC TOTAL LEAST SQUARES

We first consider the well-known case of total least squares in static models. Although it may be somewhat artificial in this case to consider observations in l_2 i.e., on an infinite time interval, it gives a better introduction for the dynamic case.

Static total least squares involves the approximation of a given time series by a regular one that satisfies linear nondynamic relations. For a required number of independent equations, the objective is to keep the approximation error

as small as possible, i.e. to minimize $\|u\|$ where u denotes the observed time series and w the approximation. If the required number of independent equations is p then the regularity of the approximation in l_2 is characterized by $Rw = 0$ for some matrix R of rank p . This means that the approximation w has rank at most $m - q - p$. In Section II (2), this class of regular time series is given by $\mathcal{B}^{(m-q-p) \times 1}$ with $n = 0$ corresponding to the exclusion of dynamic equations. This leads to the following formulation of static total least squares.

Definition 2.1 (Static Total Least Squares) For a given time series $u \in l_2^q$ and minimal required number of independent relations $p = q - m$ find a decomposition $u = w + u$ with $w \in \mathcal{B}^{q \times m \times 0}$ and with $\|u\|$ minimal.

The solution to this problem is given by the singular value decomposition (SVD). We denote the usual Euclidean norm on \mathbb{R}^q by $\|\cdot\|$ and the induced norm on l_2^q by $\|\cdot\|$.

Proposition 2.2 (Singular Value Decomposition in l_2) Every $u \in l_2^q$ can be decomposed as $u = \sum_{i=1}^q \lambda_i u_i v_i^T$ with

- 1) $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q > 0$ called the singular values of u .
- 2) $u_i \in \mathbb{R}^{1 \times 1}$ with $\|u_i\| = 1$ and $u_i^T u_j = 0$ for $i \neq j$ called the left singular vectors of u .
- 3) $v_i \in l_2^m$ with $\|v_i\| = 1$ and $v_i \perp v_j$ for $i \neq j$ called the right singular vectors of u .

The singular values are uniquely determined and if they are distinct then the singular vectors are also uniquely determined up to a sign.

Proof. The existence and properties of the SVD for finite matrices are discussed in e.g. [3]. Let Π denote the empirical covariance matrix of u i.e. $\Pi = \sum_{t=1}^n u(t)u(t)^T$. As Π is symmetric the left singular vectors are equal to the right ones so let its SVD be given by $\Pi = \sum_{i=1}^q \mu_i u_i u_i^T$. Let $U = [u_1 \dots u_q]$ then the empirical covariance matrix of $U^T u$ is given by $\text{diag}(\mu_1 \dots \mu_q)$ so $U^T u$ consists of q orthogonal components of norm $\sqrt{\mu_i}$. It is easily verified that for $v_i = (1/\sqrt{\mu_i})u_i^T u$ and $\lambda_i = \sqrt{\mu_i}$ $u = \sum_{i=1}^q \lambda_i u_i v_i^T$ is an SVD of u . \square

The SVD solves the static total least squares problem as follows.

Proposition 2.3 (Optimal Static Approximation) Let $u = \sum_{i=1}^q \lambda_i u_i v_i^T$ be the SVD of $u \in l_2^q$. Then $w = \sum_{i=1}^m \lambda_i u_i v_i^T$ is the optimal approximation of u in $\mathcal{B}^{q \times m \times 0}$ satisfying the static equations $u^T w = 0$ for $i = m+1, \dots, q$.

Proof. This result follows immediately from a corresponding property of the SVD for finite matrices cf. [3]. \square

Example. One of the essential features of total least squares is that all variables are treated in a similar way. For simplicity we consider simulated data from a model with this type of symmetry, namely an errors in variables model. This also gives us the opportunity to relate total least squares to other well-known identification methods and to discuss the role of stochastic assumptions. We consider the model

$$u(t) = \alpha(t) + \eta(t) \quad y(t) = \alpha(t) + \epsilon(t) \quad (3)$$

The observed variables consist of $u = (u \ y)$ and α is an unobserved latent variable and ϵ and η are unobserved disturbances. Three of the possible methods for the estimation of the parameter α are regression of y on u , regression of u

on y , and total least squares. These methods can be interpreted as maximum likelihood methods, in the case that ε and η are independent white noise processes with variances σ_ε^2 and σ_η^2 and if in addition respectively $\sigma_\eta^2 = 0$, $\sigma_\varepsilon^2 = 0$, or $\sigma_\varepsilon^2 = \sigma_\eta^2$.

As an example, we consider data generated by the model (3) with $\alpha = 1$ and $\sigma_\varepsilon^2 = \sigma_\eta^2 = 0.5$ and where z is a sample of 20 observations of a white noise process with variance one and independent of η and ε . The SVD of the resulting observation is

$$w = 5.64 \begin{pmatrix} 0.72 \\ 0.69 \end{pmatrix} v_1 + 2.31 \begin{pmatrix} -0.69 \\ 0.72 \end{pmatrix} v_2 \quad (4)$$

where v_1 and v_2 are two orthogonal vectors of unit length. According to Proposition 2.3, the optimal static approximation \hat{w} of w of rank 1 is given by the first term in (4), and this satisfies $[-0.69, 0.72]\hat{w} = 0$. This corresponds to an estimated value $0.69/0.72 = 0.96$, which is close to $\alpha = 1$. The corresponding approximation error is $\|w - \hat{w}\| = 2.31$. Regression of y on u yields an estimate of 0.69, and regression of u on y gives 1.36. These results are depicted in Fig. 1.

This example illustrates the fact that the static total least squares (TLS) scheme is in between both regressions. The method is easily adapted for the case $\sigma_\varepsilon^2 \neq \sigma_\eta^2$, by using a weighted l_2 -norm for $\tilde{w} = (\tilde{u}, \tilde{y})$, defined by $\|\tilde{w}\|_\alpha^2 := \alpha^2 \|\tilde{u}\|^2 + \|\tilde{y}\|^2$ with $\alpha = \sigma_\varepsilon/\sigma_\eta$. The regressions correspond to the extremal cases with infinite weight on one of the components. To choose an appropriate weighted l_2 -norm we need information on the relative errors involved in the measured system variables, c.q., the relative weight one attaches to deviations in the different variables. In this paper we will not further address these problems. \diamond

III. GLOBAL TOTAL LEAST SQUARES

In the rest of this paper we are concerned with total least squares in dynamical systems. For this purpose we describe in this section our systems concept and corresponding notions of complexity and misfit.

As stated in the introduction, we consider systems described by difference equations. From the formulation of the global total least squares problem (2), it is obvious that sets of equations with the same solution set are equivalent as they yield the same approximation error for every observation. Hence not the equations themselves, but their solution set is the essential object in our modeling procedure. This is a strong motivation for adopting the behavioral approach to systems, as introduced in [10] and [11], in which a system is defined by the set of time series that are compatible with the system laws. This set is called the behavior of a system. The system laws themselves are considered as a description or representation of the behavior. The properties we impose on difference equations are reflected by the set-theoretic properties of the corresponding behavior. Linear, time-invariant difference equations correspond to linear, shift-invariant behaviors. If in addition the equations have finite lag, it can be decided if a time series belongs to the behavior by scanning it through a finite window. This property is called completeness, which is further explained in Appendix A, (see Definition A.1). We will further restrict the attention

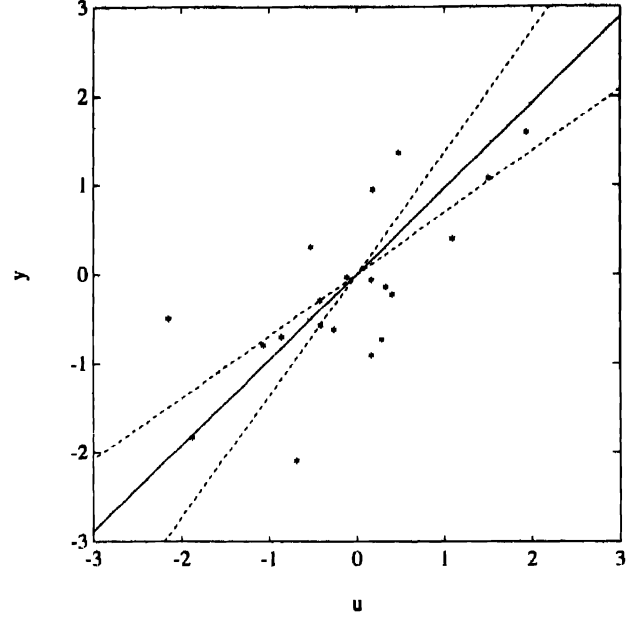


Fig. 1. Scatter diagram of the data w : the solid line denotes the TLS model and the dashed lines the two regression models; the dotted lines indicate which distance is minimized in the three different models.

to behaviors in l_2 . The resulting class of systems is defined as follows.

Definition 3.1 (l_2 -Systems): l_2 -systems are linear, shift-invariant, complete subspaces of l_2^q .

As a measure of the restrictiveness of a set of difference equations we take into account the number of independent equations and their total lag, as described in the introduction. We call the equations independent if the number of equations cannot be reduced without changing the solution set, i.e., if there exists no equivalent smaller set of equations describing the same system. On the set theoretic level of l_2 -systems, we define the complexity of a system in terms of its rank and its degree. The rank of a system is defined as the number of the degrees of freedom at each time instant, which is equal to the number of inputs, and the degree of a system corresponds to the dimension of the state space (see Definition A.3).

Definition 3.2 (Complexity): The complexity of an l_2 -system \mathcal{B} is defined as $c(\mathcal{B}) := (m, n)$, where m denotes the rank of \mathcal{B} and n its degree.

Let $\mathcal{B}^{q,m,n}$ denote the class of l_2 -systems with rank at most m and degree at most n . Then $\mathcal{B}^{q,m,n}$ consists of the solution sets of at least $q-m$ independent linear time-invariant difference equations with the sum of their lags at most n , i.e., $w \in \mathcal{B}^{q,m,n}$ if and only if there exists a system $\mathcal{B} \in \mathcal{B}^{q,m,n}$ with $w \in \mathcal{B}$.

We summarize this in Table I. For further clarification we also relate these concepts to the classical characterization of systems in terms of input-output mappings.

We define the following concept of the misfit of a system.

Definition 3.3 (Misfit): The misfit of an l_2 -system \mathcal{B} with respect to an l_2 -time series w is defined as $d(w, \mathcal{B}) := \inf_{\tilde{w} \in \mathcal{B}} \|w - \tilde{w}\|$.

This leads to the following reformulation of the global total least squares problem (GTLS) as described in the introduction (see (2)).

TABLE I
COMPARISON OF SYSTEM CONCEPTS

Difference Equations	l_2 -behaviours	I/O-mapping
linear	linear subspace	linear
time-invariant	shift-invariant	time-invariant
finite order	l_2 -complete	rational transfer function
$q-m$ indep. equations	rank m	m inputs, $q-m$ outputs
minimal sum of lags is n	degree n	McMillan degree n

Definition 3.4 (GTLS): For an observation $w \in l_2^q$ and given tolerated complexity (m, n) , determine an l_2^q -system $B^* \in \mathcal{B}^{q,m,n}$ such that $d(w, B^*) = \min_{B \in \mathcal{B}^{q,m,n}} d(w, B)$.

This involves a double minimization. The inner minimization, evaluating the misfit $d(w, B)$, amounts to optimization over a linear space. Secondly, we have to determine a system for which the misfit is minimal. This is a nonlinear optimization problem over a nonconvex set.

Leading Example: We describe a simple example, which will be used in the following sections to clarify the introduced general framework. We consider a time series in $B^{2,1,1}$, that is corrupted by white noise. The regular part w_r consists of two components, u_r and y_r , where u_r is the realization of a white noise process with unit variance, and y_r satisfies

$$y_r(t) = 2/3 y_r(t-1) + 2u_r(t) - 2u_r(t-1). \quad (5)$$

The observation w consists of two components u and y , with

$$u(t) = u_r(t) + \eta(t); \quad y(t) = y_r(t) + \varepsilon(t) \quad (6)$$

where η and ε are independent white noise processes with variance 0.25, both independent of u_r . The data consists of a time series of length 100 which is generated by system (6). To obtain a time series in l_2 , the observation (u, y) is taken to be zero outside the observation interval.

Of course, this simple example could be solved by brute force as a nonlinear parameter optimization problem, disregarding any system theoretic interpretation of the problem. In more complicated cases, however, this becomes hardly feasible. Therefore we will follow a system theoretic approach, that also gives more insight in the problem.

The GTLS results in the following sections will be compared with those obtained by three other methods, namely regression, the "output error" method, and the "local total least squares" method. Here we mention that the procedures for GTLS and local total least squares have been implemented in Matlab and that for the regression and output error method we used, respectively, the procedures ARX and OE of the System Identification Toolbox. The regression model B_{regr} is obtained by regressing $y(t)$ on $y(t-1)$, $u(t)$ and $u(t-1)$. For the observation w this gives

$$B_{\text{regr}} = \{(u, y) \in l_2^2; y(t) = 0.10y(t-1) + 1.42u(t) - 0.78u(t-1)\}. \quad (7)$$

This equation yields optimal one-step ahead predictions for $y(t)$, given $y(t-1)$, $u(t)$, and $u(t-1)$. According to our terminology, we would call this method "local ordinary least squares." By ordinary we mean that only one of the components of w is approximated and by local that only

the first step ahead predictions of difference equations are taken into account. In the GTLS scheme we approximate both components and take full account of the global, higher order forward and backward implications of difference equations.

The output error model B_{oe} is the system with the property that for the given input u , the corresponding system output y is as close as possible to the observed output y . In fact, this method has some similarity to GTLS, the difference being that the input is kept fixed and only the output is approximated. In our terminology it is a "global ordinary least squares" method. The estimated output error model is

$$B_{\text{oe}} = \{(u, y) \in l_2^2; y(t) = 0.68y(t-1) + 1.15u(t) - 1.51u(t-1)\}. \quad (8)$$

Finally we use a simple modification of the static total least squares method of Section II to determine a first-order model, as follows. Define the block Hankel matrix $H \in l_2^4$ by $H(t) = \begin{bmatrix} w(t-1) \\ w(t) \end{bmatrix}$, then static equations for H correspond to first-order equations for the observation w . The optimal static equation for H is obtained by applying Proposition 2.2 with $q = 4$ and $m = 3$. Clearly, the quality of the corresponding first-order model for w is evaluated only locally. By this we mean that, for example, the second-order restrictions on $(w(t-2), w(t-1), w(t))^T$ implied by the model are not taken into account, and the same holds true for higher order restrictions. Therefore we call this the "local total least squares" model. For the data of this example this gives

$$B_{\text{lls}} = \{(u, y) \in l_2^2; y(t) = 0.54y(t-1) + 1.87u(t) - 1.82u(t-1)\}. \quad (9)$$

This example will be continued in the next sections. \diamond

IV. STATE REPRESENTATIONS

One of the crucial questions in our modeling theory is how to calculate the misfit of a system with respect to a given observation, cf. Definition 3.3. Obviously this requires a numerical representation of the system. In Section V we develop a representation that is extremely useful for this purpose, namely the isometric state representations. They also play a central role in the construction of optimal models, as discussed in Section VII. We now first introduce general state representations, which will be abbreviated as SR. Let σ denote the shift operator, defined as $\sigma x(t) = x(t+1)$.

Definition 4.1 (State Representation). A state representation (A, B, C, D) of an l_2 -system B is a description of the form

$$B = \{w \in l_2^q; \exists x \in l_2^n, v \in l_2^m \text{ such that } \sigma x = Ax + Bv \text{ and } w = Cx + Dv\}$$

with $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{q \times n}$, $D \in \mathbb{R}^{q \times m}$ and $n, m \in \mathbb{N}$.

Here v is an auxiliary input, x is a state trajectory and w a system trajectory, m denotes the number of auxiliary inputs and n the number of state variables. The system defined by

this representation is denoted as $\mathcal{B}(A, B, C, D)$. We mention that the condition that x and v are square summable involves no loss of generality, as it can be shown that for every $(x, v): \mathbf{Z} \rightarrow \mathbf{R}^n \times \mathbf{R}^m$ generating $w \in l_2^q$ there also exists $(x', v') \in l_2^n \times l_2^m$ generating the same time series w . Systems in $\mathbf{B}^{q,m,n}$, i.e., with rank at most m and degree at most n , are precisely those systems that admit an SR with m auxiliary inputs and n states, cf. Proposition A.4.

Representations are called equivalent if they describe the same system. They are called minimal if the number of state variables and the number of auxiliary inputs are both minimal. These two quantities can be minimized simultaneously, so for every system there indeed exists a minimal SR. From a given SR we obtain equivalent ones as follows.

Proposition 4.2 (Equivalent State Representations): Let be given a state representation (A, B, C, D) of an l_2 -system \mathcal{B} . Then for all invertible $S \in \mathbf{R}^{n \times n}$, invertible $R \in \mathbf{R}^{m \times m}$, and $F \in \mathbf{R}^{m \times n}$, $(S(A + BF)S^{-1}, SBR, (C + DF)S^{-1}, DR)$ is also a state representation of \mathcal{B} . Moreover, if (A, B, C, D) is minimal, then all minimal state representations for \mathcal{B} are obtained in this way.

Proof: By definition, $w \in l_2^q$ is contained in $\mathcal{B}(A, B, C, D)$ if and only if there exist an v and x such that $\sigma x = Ax + Bv$ and $w = Cx + Dv$. These equations are equivalent to $\sigma(Sx) = S(A + BF)S^{-1}(Sx) + SBR(R^{-1}(v - Fx))$ and $w = (C + DF)S^{-1}(Sx) + DR(R^{-1}(v - Fx))$, provided that S and R are invertible. This shows the equivalence of the representations. For a proof of the fact that for minimal representations all equivalent representations are obtained in this way we refer to [5, Corollary IV.3-4]. \square

Observe that minimal SR's for a given l_2 -system are highly nonunique. The choice of basis in the state space, corresponding to S , is a well-known nonuniqueness of state-space representations. In our framework the auxiliary input v is merely a tool to describe the system behavior and need not have additional external significance. This allows for a basis transformation for the auxiliary input, corresponding to R . Further the behavior is invariant under a static state feedback F to this auxiliary input. This is in contrast to the common notion of feedback to the actual input of the system, which would affect the set of compatible input-output pairs. As a consequence, in our framework the spectrum of the A -matrix is not an intrinsic property of a system. In the next section we exploit this nonuniqueness to obtain SR's with convenient properties for computing the misfit of a model with respect to data.

Minimality of a representation can be expressed in terms of rank conditions on the matrices (A, B, C, D) , as follows.

Proposition 4.3 (Minimal State Representations for l_2 -Systems): A state representation (A, B, C, D) is minimal if and only if:

- 1) (A, B) is controllable
- 2) $\forall F \in \mathbf{R}^{m \times n}$, $(A + BF, C + DF)$ is observable
- 3) $\ker D = 0$.

Proof: See Appendix B.

By eliminating the state, SR's induce an image representation of l_2 -systems. For the exposition in the sequel,

representations with A asymptotically stable are most relevant. Every l_2 -system allows for such a representation, which follows from Propositions 4.2 and 4.3.1.

Proposition 4.4 (Image Representations): Let \mathcal{B} be an l_2^q -system in $\mathbf{B}^{q,m,n}$, with state representation (A, B, C, D) , where A is asymptotically stable. Let $G: l_2^m \rightarrow l_2^q$ be defined as $G = C(\sigma I - A)^{-1}B + D$, i.e., $(Gv)(t) = Dv(t) + \sum_{k=1}^{\infty} CA^{k-1}Bv(t-k)$. Then $\mathcal{B} = \text{im } G$.

Proof: G is well defined, as A is asymptotically stable. It is easily verified that the equations $\sigma x = Ax + Bv$ and $w = Cx + Dv$ imply that $Gv = w$. \square

Leading Example (Continued): Consider the l_2 -system corresponding to (5), $\mathcal{B}_{\text{ex}} := \{w \in l_2^2; w = (u, y), \text{ with } y(t) = 2/3y(t-1) + 2u(t) - 2u(t-1)\}$. This can be written in input/state/output form as follows

$$x(t+1) = 2/3x(t) + u(t); y(t) = -2/3x(t) + 2u(t). \quad (10)$$

From this it is easy to obtain an SR by taking the auxiliary input v to be equal to u , which gives

$$\mathcal{B}_{\text{ex}} = \mathcal{B}\left(2/3, 1, \begin{bmatrix} 0 \\ -2/3 \end{bmatrix}\right) \quad (11)$$

This representation is minimal. The corresponding image representation $G = \sum_{k=0}^{\infty} G_k \sigma^{-k}$ has coefficients $G_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, and $G_k = \begin{bmatrix} 0 \\ -(2/3)^k \end{bmatrix}$. An SR equivalent to (11) is given by $\left(1, 1/2, \begin{bmatrix} 1/3 \\ 0 \end{bmatrix}, \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}\right)$, which follows from Proposition 4.2 by taking $S = 1$, $R = 1/2$, and $F = 1/3$. Note that in this representation the auxiliary input is given by y . This also illustrates that the eigenvalues of A are not intrinsic for the system \mathcal{B} . \diamond

V. ISOMETRIC STATE REPRESENTATIONS

In this section we define isometric state representations (ISR's), which are defined by a local isometry property involving the state variable.

Definition 5.1 (Isometric State Representation): A state representation (A, B, C, D) is called isometric if for all $x \in \mathbf{R}^n$, $v \in \mathbf{R}^m$, $w \in \mathbf{R}^q$ and $z \in \mathbf{R}^n$ such that $z = Ax + Bv$ and $w = Cx + Dv$ there holds

$$|v|^2 + |x|^2 = |w|^2 + |z|^2. \quad (12)$$

Equivalently

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^T \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I_n & 0 \\ 0 & I_m \end{pmatrix} \quad (13)$$

Minimal ISR's can be constructed from arbitrary minimal SR's as follows.

Proposition 5.2 (Construction of ISR): Let (A, B, C, D) be a minimal state representation, and let $K \in \mathbf{R}^{n \times n}$ be the unique symmetric positive definite solution of the algebraic Riccati equation

$$K = A^T K A - (B^T K A + D^T C)^T (B^T K B + D^T D)^{-1} \cdot (B^T K A + D^T C) + C^T C. \quad (14)$$

Let the matrices $S \in \mathbf{R}^{n \times n}$, $F \in \mathbf{R}^{m \times n}$ and $R \in \mathbf{R}^{m \times m}$ be solutions of the equations

$$S^T S = K \quad (15)$$

$$RR^T = (B^T K B + D^T D)^{-1} \quad (16)$$

$$F = -(B^T K B + D^T D)^{-1}(B^T K A + D^T C). \quad (17)$$

Then $(S(A + BF)S^{-1}, SBR, (C + DF)S^{-1}, DR)$ is an equivalent isometric minimal state representation.

Proof: See Appendix B.

The following proposition gives a necessary minimality condition for ISR's and states that they are unique modulo unitary transformations.

Proposition 5.3 (Minimal ISR).

- 1) If (A, B, C, D) is an ISR, then A is stable. If the representation is minimal then A is asymptotically stable.
- 2) Two minimal ISR's (A, B, C, D) and (A', B', C', D') are equivalent if and only if there exist unitary matrices U and V such that $(A', B', C', D') = (UAU^T, UBV, C'U^T, DV)$.

Proof: See Appendix B.

ISR's induce a description of l_2 -systems as the image of an isometric operator. This is made explicit in the following proposition.

Proposition 5.4 (Isometric Image Representations): Let (A, B, C, D) be a minimal isometric state representation of an l_2 -system \mathcal{B} . Then the corresponding image operator $G: l_2^m \rightarrow l_2^q$ defined by $G = C(\sigma I - A)^{-1}B + D$ is isometric, i.e., $\|Gv\| = \|v\|$ for all $v \in l_2^m$.

Proof: If $w = Gv$ then $w(t) = Cx(t) + Dv(t)$ for $x = (\sigma I - A)^{-1}Bv$, so that $\sigma x = Ax + Bv$. Summation of (12) over $t \in \mathbf{Z}$ yields $\|v\|^2 + \|w\|^2 + \|\sigma x\|^2$. Clearly $\|\sigma x\| = \|\sigma x\|$, from which the result follows. \square

Summarizing, every l_2 -system can be represented as the image of an isometric $q \times m$ transfer function with an ISR as its realization. In the literature an isometric operator G is sometimes called lossless, and if it is in addition stable it is called inner. So an ISR is a realization of an inner transfer function that displays the isometry in a local way, in terms of the state variables.

Leading Example (Continued): We apply the construction of Proposition 5.2 to the SR (11) of l_2 -system \mathcal{B}_{ex} . This yields $K = 4/9$, $S = 2/3$, $R = 3/7$, and $F = 4/21$, resulting in the isometric representation

$$\sigma x = 6/7x + 2/7v; w = \begin{bmatrix} 2/7 \\ -3/7 \end{bmatrix} x + \begin{bmatrix} 3/7 \\ 6/7 \end{bmatrix} v. \quad (18)$$

According to Proposition 5.3, \mathcal{B}_{ex} has a unique minimal ISR, modulo sign changes for the state and for the auxiliary input. \diamond

VI. OPTIMAL APPROXIMATION WITHIN A SYSTEM

In this and the next section we consider the GTLS problem of Definition 3.4. As stated before, this involves a double minimization. In this section we discuss the computation of

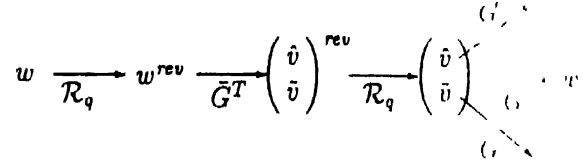


Fig. 2. Projection scheme.

the misfit, and in the next section we treat the problem of selecting an optimal model.

For the computation of the misfit we will use the adjoint G^* of an operator $G: l_2^m \rightarrow l_2^q$. This is defined by the condition $\langle w, Gv \rangle = \langle G^*w, v \rangle$ for all $v \in l_2^m$ and $w \in l_2^q$, where $\langle \cdot, \cdot \rangle$ denotes the inner product on l_2 . It follows that G is isometric if and only if $G^*G = I$. For the image operator $G := C(\sigma I - A)^{-1}B + D$ the adjoint is given by $G^* = B^T(\sigma^{-1}I - A^T)^{-1}C^T + D^T$. This can also be written as $G^* = \mathcal{R}_m G^T \mathcal{R}_q$ where $\mathcal{R}_k: l_2^k \rightarrow l_2^k$ denotes the time reversion operator defined by $\mathcal{R}_k w(t) := w(-t)$ and $G^T := B^T(\sigma I - A^T)^{-1}C^T + D^T$.

Theorem 6.1 (Optimal Approximation Within a System): Let be given an observation $w \in l_2^q$ and an l_2^q -system $\mathcal{B} \in \mathbf{B}^{q,m,n}$. Let $G: l_2^m \rightarrow l_2^q$ be an isometric image representation of \mathcal{B} . Then the optimal approximation $\hat{w} \in \mathcal{B}$ is given by $\hat{w} = GG^*w$, with misfit $d(w, \mathcal{B}) = \|(I_q - GG^*)w\|$.

Proof: This is a well-known result. To be explicit, we should minimize $\|w - Gv\|$ over $v \in l_2^m$. Let $v' = v - G^*w$, then $\|w - Gv\|^2 = \|w - GG^*w - Gv'\|^2 = \|GG^*w\|^2 + \|Gv'\|^2$, as $\langle w - GG^*w, Gv' \rangle = \langle G^*w - G^*w, v' \rangle = 0$ where we use that G is isometric so that $G^*G = I_m$. So the minimum is achieved by taking $v' = 0$, and $w = GG^*w$. \square

The optimal approximation within a system \mathcal{B} gives rise to a decomposition $w = \hat{w} + \tilde{w}$, with $\hat{w} \in \mathcal{B}$ regular and with \tilde{w} the corresponding approximation error. We will now show that \tilde{w} also exhibits regularity. As \hat{w} is obtained as the projection of w on \mathcal{B} , it follows that $\tilde{w} \in \mathcal{B}^\perp = \{w \in l_2; \langle w, w' \rangle = 0 \text{ for all } w' \in \mathcal{B}\}$. This set is clearly linear and shift-invariant. The following theorem states that it is an l_2 -system.

Theorem 6.2 (Orthogonal Complement). Let \mathcal{B} be an l_2 -system with rank m and degree n .

- 1) The orthogonal complement \mathcal{B}^\perp of \mathcal{B} is an l_2 -system.
- 2) $\mathcal{B} \oplus \mathcal{B}^\perp = l_2^q$.
- 3) \mathcal{B}^\perp has rank $q - m$ and degree n .
- 4) Let (A, B, C, D) be a minimal isometric state representation of \mathcal{B} , and let \tilde{B}, \tilde{D} be such that $\begin{pmatrix} A & B & \tilde{B} \\ C & D & \tilde{D} \end{pmatrix}$ is a unitary matrix. Then $(A, \tilde{B}, C, \tilde{D})$ is a minimal isometric state representation of \mathcal{B}^\perp .
- 5) Let \hat{w} be the optimal approximation in \mathcal{B} of $w \in l_2^q$. Then the approximation error $\tilde{w} := w - \hat{w}$ is the optimal approximation in \mathcal{B}^\perp of w .

Proof: See Appendix B.

We summarize this result in the following projection scheme as shown in Fig. 2. Here G and \tilde{G} denote the image representations corresponding to, respectively, (A, B, C, D) and $(A, \tilde{B}, C, \tilde{D})$, and $\bar{G} := [G \ \tilde{G}]$. In the literature \bar{G} is sometimes called a lossless embedding of G . Further \hat{v} and \tilde{v} are

TABLE II
MISFIT OF MODELS

model	\mathcal{B}_{ex}	\mathcal{B}_{oe}	\mathcal{B}_{regt}	\mathcal{B}_{ltls}	\mathcal{B}_{rand}
misfit	5.22	5.62	6.84	5.30	13.53

the auxiliary inputs corresponding to \hat{w} and \tilde{w} , i.e., $\hat{w} = G\hat{v}$ and $\tilde{w} = \tilde{G}\tilde{v}$. According to Theorem 6.1 and 6.2.5, there holds that $\hat{v} = G^*w$ and $\tilde{v} = \tilde{G}^*w$ as shown in Fig. 2.

Leading Example (Continued): We apply the projection algorithm of Theorem 6.1 to determine the optimal approximation $\hat{w} \in \mathcal{B}_{ex}$ of the observation w described in Section III. Let G denote the image operator corresponding to the ISR (18), and let (A, B, C, D) denote the corresponding matrices. First compute $\hat{v} := G^*w$, which is given by the backward state equations $x(t) = A^T x(t+1) + C^T w(t)$; $\hat{v}(t) = B^T x(t+1) + D^T w(t)$. Then $\hat{w} = G\hat{v}$ is given by $\hat{x}(t+1) = A\hat{x}(t) + B\hat{v}(t)$; $\hat{w}(t) = C\hat{x}(t) + D\hat{v}(t)$.

According to Theorem 6.2, \mathcal{B}_{ex}^\perp has ISR $(6/7, 3/7, \begin{bmatrix} 2/7 \\ -3/7 \end{bmatrix}, \begin{bmatrix} -6/7 \\ 2/7 \end{bmatrix})$. This corresponds to the equation

$$y(t) = y(t-1) - 1/3u(t) + 1/2u(t-1). \quad (19)$$

So every time series in l_2^2 that is orthogonal to \mathcal{B}_{ex} satisfies this difference equation, in particular the approximation error $\tilde{w} := w - \hat{w}$. The misfit of \mathcal{B}_{ex} equals $\|\tilde{w}\| = 5.22$, as compared to $\|w\| = 22.72$. This is considerably smaller than the l_2 -norm of the white noise by which the observation was corrupted. Recall that $w = w_r + w_n$, with $w_r \in \mathcal{B}_{ex}$ and w_n white noise. In our example, $\|w_n\| = 7.62$. The optimal approximation is simply obtained by projecting the noise w_n on \mathcal{B}_{ex} , with a resulting decomposition $w_n = \hat{w}_n + \tilde{w}_n$, where $\hat{w}_n \in \mathcal{B}_{ex}$ and $\tilde{w}_n \in \mathcal{B}_{ex}^\perp$. In our case, $\|\hat{w}_n\| = 5.56$ and $\|\tilde{w}_n\| = \|\tilde{w}\| = 5.22$. We also determine projections onto the models \mathcal{B}_{regt} , \mathcal{B}_{oe} , and \mathcal{B}_{ltls} , as defined in Section III, and onto a randomly chosen system $\mathcal{B}_{rand} \in \mathcal{B}^{2,1,1}$. The parameters of an SR of \mathcal{B}_{rand} were obtained by a random sample from the standard normal distribution. The resulting misfits are listed in Table II. It turns out that the local total least squares model is of relatively good quality in this example. This, however, may be completely different in other situations, as we will illustrate by an example in Section IX-B. \diamond

VII. MODEL IMPROVEMENT CONSTRUCTIONS

In this section we discuss the second part of the GTLS problem, namely determining an l_2 -system with minimal misfit with respect to a given observation. Formulated in terms of state representations, this amounts to the following.

Definition 7.1 (GTLS in terms of SR): For given observation $w \in l_2^q$ and tolerated complexity (m, n) , determine a state representation (A, B, C, D) with m auxiliary inputs and n states and an auxiliary input $\hat{v} \in l_2^m$ such that for $\hat{w}(t) := (C(\sigma I - A)^{-1}B + D)\hat{v}$, the error $\|w - \hat{w}\|$ is minimal.

We follow an iterative approach for this nonlinear problem. In each step we keep some parameters fixed, such that the resulting subproblem becomes sufficiently simple. For instance, for fixed (A, B, C, D) the resulting problem in \hat{v} is solved

by the projection scheme discussed in the foregoing section. We consider the following subproblems.

Problem 1—Optimal C and D : For given A, B and \hat{v} , solve the GTLS problem for C and D .

Problem 2—Optimal B and D : For given A, C and \hat{v} , solve the GTLS problem for B and D .

Problem 3—Optimal B, D and \hat{v} in an ISR: For given A and C with $A^T A + C^T C = I_n$, solve the GTLS problem for B, D , and \hat{v} , under the restriction that (A, B, C, D) is isometric.

In the next theorem we give constructive solutions of these problems.

Theorem 7.2 (Model Improvement Constructions):

- 1) Construction 1 (Projection of the Approximation Error): Let \hat{x} be defined by $\sigma\hat{x} = A\hat{x} + B\hat{v}$, and let $\mathcal{E} := \{\hat{w} \in l_2^q; \exists C \in \mathbb{R}^{q \times n}, D \in \mathbb{R}^{q \times m} \text{ such that } \hat{w} = C\hat{x} + D\hat{v}\}$. Let $P\hat{x} + Q\hat{v}$ denote the orthogonal projection of w onto \mathcal{E} , then $\mathcal{B}_1 := \mathcal{B}(A, B, P, Q)$ solves problem 1.
- 2) Construction 2 (Dual version of construction 1): Let $\mathcal{F} := \{\hat{w} \in l_2^q; \exists B \in \mathbb{R}^{n \times m}, D \in \mathbb{R}^{q \times m} \text{ such that } \hat{w} = (C(\sigma I - A)^{-1}B + D)\hat{v}\}$. Let the orthogonal projection of w onto \mathcal{F} be given by $(C(\sigma I - A)^{-1}P + Q)\hat{v}$, then $\mathcal{B}_2 := \mathcal{B}(A, P, C, Q)$ solves problem 2.
- 3) Construction 3 (SVD on auxiliary inputs): Let $\bar{B} \in \mathbb{R}^{n \times q}, \bar{D} \in \mathbb{R}^{q \times q}$ be such that $\begin{pmatrix} A & \bar{B} \\ C & \bar{D} \end{pmatrix}$ is a unitary matrix. Let $\bar{v} := (\bar{B}^T(\sigma^{-1}I - A^T)^{-1}C^T + \bar{D}^T)w$ have SVD $\bar{v} = \sum_{i=1}^q \lambda_i u_i z_i$, and let $U_m := [u_1, \dots, u_m]$. Then $\mathcal{B}_3 := \mathcal{B}(A, \bar{B}U_m, C, \bar{D}U_m)$ solves problem 3.

Proof: Parts 1 and 2 follow immediately from the definitions of \mathcal{E} and \mathcal{F} . For part 3, observe that (A, B, C, D) is an ISR if and only if there exist \bar{B}, \bar{D} , and a unitary V such that $[B \ \bar{B}] = \bar{B}V$ and $[D \ \bar{D}] = \bar{D}V$. For this representation the projection scheme gives $\begin{pmatrix} v \\ \bar{v} \end{pmatrix} = V^T \bar{v}$, with misfit $\|\bar{v}\|$. By taking $V = [u_1, \dots, u_q]$, this misfit is determined by the $q-m$ smallest singular values of \bar{v} , which is minimal. \square

We use these results in an iterative algorithm for the GTLS problem. At each step we use one of the three constructions to improve the model. The resulting model parameters are transformed to ISR, which also involves an update of the A -matrix. The projection scheme is then applied to update v .

Proposition 7.3: The above method leads to a sequence of models with monotonically decreasing misfit.

Proof: This is immediately evident from Theorem 7.2. \square

So the algorithm leads to a convergent sequence of misfits. In the next section we show that, in the limit, the corresponding models are stationary points with respect to the GTLS criterion.

Leading Example (Continued): To illustrate the foregoing, we consider the data $w = (u, y)$ described in Section III (see (5) and (6)). To investigate the effect of the choice of an initial model, we apply the model improvement constructions to the models \mathcal{B}_{ex} , \mathcal{B}_{oe} , \mathcal{B}_{regt} , \mathcal{B}_{ltls} , and \mathcal{B}_{rand} as described in Sections III and VI. The results are in Table III. The first row shows the initial misfits, cf. Table II. The next three rows contain the misfits of the models obtained by applying each of the constructions separately and only once. This shows that

TABLE III
MODEL IMPROVEMENTS

model	B_{ex}	B_{oe}	B_{regr}	B_{ltls}	B_{rand}
misfit	5.22	5.62	6.84	5.30	13.53
construction 1	5.14	5.12	6.10	5.23	6.21
construction 2	5.14	5.13	6.13	5.30	6.26
construction 3	5.21	5.15	6.78	5.30	6.59
limit	5.11	5.11	5.11	5.11	5.11

TABLE IV
ORDER SELECTION

order	$n = 0$	$n = 1$	$n = 2$	$n = 3$	$n = 4$
misfit	6.71	5.11	5.02	4.91	4.91

each individual construction can give a significant decrease of the misfit. The last row shows the misfit resulting from applying these constructions iteratively until convergence.

For each initial model convergence occurred after about 20 iterations. The limiting model is the same in all four cases, which suggests that it is optimal. It is given by

$$B_{\text{GTLs}} = \{(u, y) \in l_2^2; y(t) = 0.67y(t-1) + 1.85u(t) - 1.96u(t-1)\}. \quad (20)$$

The parameters of this system are relatively close to those of the data generating system, cf. (5).

Next, we investigate whether the model order can be deduced from the data. For this purpose we compare in Table IV the optimal misfits for models of various degree. The misfit of the optimal static model is given by the smallest singular value of w . This clearly motivates the choice of a first-order model. It is significantly better than the static model, and an increase of the order gives only small improvements. \diamond

VIII. OPTIMALITY CONDITIONS

From the model improvement constructions in Section VII we can derive necessary conditions for optimality, as for an optimal model the constructions can give no improvement. We express the optimality conditions in terms of empirical covariances. For two sequences $a \in l_2^k$ and $b \in l_2^l$ this is defined as $\text{cov}(a, b) := \sum_{t=-\infty}^{\infty} a(t)b(t)^T \in \mathbb{R}^{k \times l}$. Further, by $\text{cov}([a_1, a_2], [b_1, b_2])$, we denote the covariance matrix of the combined trajectories $[a_1^T \ a_2^T]^T$ and $[b_1^T \ b_2^T]^T$.

Theorem 8.1 (l_2 -optimality Conditions): Let B denote a GTLS model for an observation $w \in l_2^q$. Let $\hat{w} \in B$ denote the optimal approximation of w , and $\tilde{w} \in B^\perp$ the corresponding approximation error. Let \hat{x}, \hat{v} denote, respectively, the state and auxiliary input corresponding to \hat{w} in a minimal state representation of B , and let \tilde{x}, \tilde{v} be defined analogously for \tilde{w} . Then the following equivalent conditions hold:

- 1) $\text{cov}(\hat{v}, \tilde{v}) = 0$, $\text{cov}(\hat{v}, \hat{x}) = 0$ and $\text{cov}(\hat{x}, \tilde{v}) = 0$;
- 2) $\text{cov}([\hat{v}, \hat{x}], [\tilde{v}, \tilde{x}]) = 0$;
- 3) $\text{cov}([\hat{v}, \hat{x}, \hat{w}, \sigma x], [\tilde{v}, \tilde{x}, \tilde{w}, \sigma \tilde{x}]) = 0$.

Proof: See Appendix B.

In practice, to evaluate how far these conditions are satisfied it may be useful to consider the empirical correlations, i.e., the covariances scaled by the magnitude of the variables.

Next we investigate how far these conditions are sufficient for optimality. It is not difficult to check that the number of free parameters in (A, B, C, D) , modulo the equivalence of Proposition 4.2, is given by $nq + m(q - m)$. This is precisely the number of equations in Theorem 8.1.1. In fact, these conditions characterize the stationary points with respect to the GTLS criterion. We call a system B a stationary point for an observation w if all the derivatives of the GTLS misfit $d(w, B(A, B, C, D))$ with respect to the system parameters are zero for a minimal SR of B . As all its minimal SR's are linearly related this is equivalent to the condition that all minimal SR's of B are stationary points.

Theorem 8.2: An l_2 -system B satisfies the optimality conditions of Theorem 8.1 if and only if B is a stationary point of the GTLS criterion.

Proof: See Appendix B.

This shows that the GTLS algorithm can only converge to stationary points. This does not, however, establish convergence of the systems. For what it is worth, we mention that we never encountered convergence problems in any of our simulations. A thorough discussion of the convergence properties of the algorithm falls beyond the scope of this paper.

The foregoing results can also be used to analyze whether a proposed system B is close to optimality. This is, for example, relevant in the formulation of stopping criteria for the iterative algorithm of Section VII. Probably the most convincing way to evaluate optimality is to consider the distance between B and a GTLS model B^* , as defined in Definition 3.4. This is in general not feasible, however, as it would require the knowledge of B^* . Instead of asking how far the system should be changed to become optimal for the observed data w , we will consider the question of how far these data should be changed to make the given system optimal. For pragmatic reasons we consider the distance to the nearest stationary point, defined as

$$\min\{\|\bar{w}\|; B \text{ is stationary for } w - \bar{w}\}. \quad (21)$$

Because it seems difficult to evaluate this distance exactly, we present an upper bound that is relatively easy to compute. This upper bound is obtained by allowing only adjustments of the data that belong to B^\perp , so that the optimal approximation of the data within B is not affected. This leads to the following definition of the optimality margin.

Definition 8.3 (Optimality Margin): The optimality margin of a system B with respect to an observation w is defined as

$$\min\{\|\bar{w}\|; \bar{w} \in B^\perp \text{ and } B \text{ is stationary for } w - \bar{w}\}. \quad (22)$$

The following result shows that the computation of the optimality margin is indeed relatively easy.

Proposition 8.4 (Optimality Margin): Let w denote the optimal approximation of w in B and let $\tilde{w} := w - \hat{w}$ denote the corresponding approximation error. Further define $\mathcal{Z} := \{z \in B^\perp; B \text{ is stationary for } \hat{w} + z\}$. Then \mathcal{Z} is a linear space, and the optimality margin is given by $\|\tilde{w} - \tilde{w}'\|$, where \tilde{w}' is the orthogonal projection of \tilde{w} on \mathcal{Z} .

TABLE V
OPTIMALITY MARGINS

model	B_{gtls}	B_{ex}	B_{oe}	B_{reg}	B_{lts}	B_{rand}
optimality margin	$4.58 \cdot 10^{-6}$	1.00	2.29	3.32	1.23	11.19

Proof: See Appendix B.

These results can also be used to determine a lower bound for the achievable misfit. This indicates the quality of a proposed model relative to the optimal one. For this purpose we will assume that a proposed model B is not only stationary for the adjusted data $w - \bar{w}$, but even globally optimal.

Proposition 8.5 (Bounds for the Minimal Misfit): Let B be given an observation w , and let e^* be the minimally achievable misfit under a certain complexity constraint. Further let B be a GTLS model of tolerated complexity for adjusted data $w - \bar{w}$ with misfit $e := d(w, B)$. Then there holds

$$e - 2\|\bar{w}\| \leq e^* \leq e. \quad (23)$$

Proof: Let B^* denote a GTLS model for the original data w , so that $d(w, B^*) = e^*$. Then the upper bound follows from the optimality of B^* . For the lower bound we use the properties of the misfit that $d(w, B) \leq \|w\|$ and $d(w_1 + w_2, B) \leq d(w_1, B) + d(w_2, B)$, so that $e = d(w, B) \leq d(\bar{w}, B) + d(w - \bar{w}, B) \leq \|\bar{w}\| + d(w - \bar{w}, B^*) \leq \|\bar{w}\| + d(w, B^*) + d(\bar{w}, B^*) \leq 2\|\bar{w}\| + e^*$. Here we have used the optimality of B for $w - \bar{w}$ in the second inequality. \square

This shows that for models with a small optimality margin the corresponding misfit is nearly optimal.

Leading Example (Continued): The covariances of Theorem 8.1 give a first indication of the optimality of a model. To make this scale invariant we consider the correlations in an ISR. For the nominal model B_{ex} they are around 0.3, for the regression model B_{reg} around 0.35, while for the randomly chosen model B_{rand} correlations of 0.8 occur. For the model B_{gtls} in (20) the correlations are approximately zero, below 10^{-5} .

From the optimality margins we obtain more precise information about the optimality of the systems. They are listed in Table V. This shows that it requires only a change of the observation of the order $\|\bar{w}\| \approx 10^{-5}$ to make B_{gtls} a stationary point, cf. (22). Now assume that B_{gtls} is globally optimal for $w - \bar{w}$, which is reasonable assumption. The evaluation of the bounds in Proposition 8.5 for the data in this example with $e = 5.1095$ and $2\|\bar{w}\| \approx 10^{-5}$ shows that the optimal misfit $e^* \approx 5.1095$ is determined within an accuracy of 10^{-5} . This also shows that B_{gtls} is optimal within this accuracy level.

IX. SIMULATION EXPERIMENTS

We illustrate the use of the GTLS algorithm by three simulation experiments. The first example concerns model reduction, i.e., the approximation of a system by one of lower complexity. We use weighted l_2 -norms to determine the l_2 -optimal approximation of a systems impulse response. In the second example we show that the algorithm can handle noncausal systems without any additional difficulty. As a final example we identify a system with multiple outputs, described

TABLE VI
PROPERTIES OF REDUCED MODELS

model	B_{gtls}	B_{bal}	B_{hank}
misfit with respect to w	0.32	0.36	0.38
error in impulse response	0.54	0.46	0.57
Hankel norm distance	0.93	0.80	0.74

TABLE VII
EFFECT OF SCALING

scaling factor	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$
$\ \cdot\ _\alpha$ -misfit with respect to w	0.3204	0.4457	0.4476
error in impulse response	0.5407	0.4477	0.4476

by more than one equation, and we discuss the choice of the model complexity.

A. l_2 -Model Reduction

The algorithm of Section VII can be applied to arbitrary time series in l_2 . Here we analyze its performance for very special data, a system impulse response. The aim is to reduce the dimension of the state space in such a way that the error in the impulse response is as small as possible (cf. [9] and the references therein). We compare the results of our algorithm with those obtained by balanced reduction and optimal Hankel norm approximation that have been developed especially for model reduction (see [2] and [7]). We consider the single-input, single-output system B with poles in $\pm 0.9i$ and $-0.7 \pm 0.6i$, so $B = \{[u \ y]^T \in l_2^2; y(t) = 0.5u(t) - 1.4y(t-1) - 1.66y(t-2) - 1.13y(t-3) - 0.69y(t-4)\}$. This system has complexity (1, 4), and we consider reduction to complexity (1, 2). The observation $w \in B$ consists of two components u and y , where u is a unit pulse at time $t = 0$ and y is the corresponding response.

We apply the GTLS algorithm, starting in a randomly chosen model. When the decrease in the misfit has become sufficiently small, below 10^{-10} , the iterations are stopped. This occurs after a few hundred iterations. The final model B_{gtls} is compared in Table VI with the balanced reduction B_{bal} and the Hankel norm reduction B_{hank} .

The l_2 -error in the impulse response in B_{gtls} is somewhat larger than that in B_{bal} and B_{hank} . If one is interested in this response then one should prevent an approximation of the input, so that an optimal approximation of the output becomes the criterion. This is achieved by taking the norm $\|\tilde{w}\|_\alpha^2 := \alpha^2 \|\tilde{u}\|^2 + \|\tilde{y}\|^2$ with α sufficiently large. The effect of increasing α is given in Table VII. This shows that for large α the method determines better approximations of the impulse response.

This also gives bounds for the minimally achievable l_2 -error, which we denote by e^* . Let B_α be the GTLS model for $\|\cdot\|_\alpha$, and let y_α be the impulse response of B_α ; then it is easily checked that $d(w, B_\alpha) \leq e^* \leq \|y - y_\alpha\|$. By increasing α we can obtain an arbitrarily accurate estimate of e^* . This gives an iterative solution method for the l_2 -optimal impulse response approximation problem. For $\alpha = 100$ we obtain $e^* = 0.4476$; see Table VII. The corresponding model

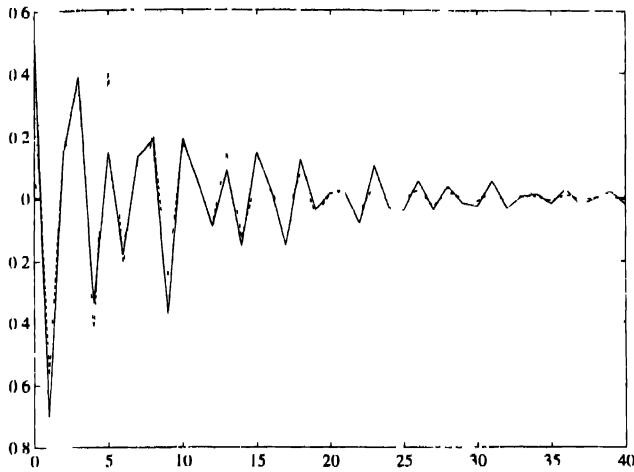


Fig. 3 The impulse response (solid line) and its l_2 -optimal approximation of second order (dashed line)

is given by the equation $y(t) = -1.35y(t-1) - 0.78y(t-2) + 0.50u(t) + 0.11u(t-1) - 0.21u(t-2)$. The impulse response is depicted in Fig. 3.

B Noncausal Systems

We consider the following noncausal system, the "Mexican hat,"

$$w_2(t) = -\frac{d^2}{dt^2} \left\{ \int_{-\infty}^{\infty} \varphi(x) w_1(t-x) dx \right\} \quad (24)$$

where φ is the standard normal density $\varphi(r) = (2\pi)^{-1/2} e^{-\frac{1}{2}r^2}$. In the simulations we consider a discrete-time version $w_2(t) = \sum_{j=-N}^N G_j w_1(t-j)$ with $N = 40$ and time steps of size 0.2. Note that w_2 is not a causal output, as the transfer function from w_1 to w_2 is not proper.

First we apply our procedure to the impulse response observation, i.e., w_1 is a unit pulse at time $t = 0$ and w_2 is the corresponding response; see Fig. 4. The misfits of the optimal models of orders 2, 4, and 6 are given in Table VIII. They are compared with the optimal Hankel norm approximations of orders 1, 2, and 3 of the causal part of the impulse response and using the symmetry of the Mexican hat to estimate the anticausal part. Analogously we determined approximations by balanced reduction. In Table VIII we also list the error in the impulse response of these models, i.e., the l_2 -distance between the systems impulse response and the Mexican hat w_2 . This error should be compared with the magnitude of the response, given by $\|w_2\| = 0.35$.

Hankel norm reduction and especially balancing give rather good results. They can only be used, however, when a causal impulse response is available. The GTLS method makes no use of the symmetry of the observed signals, but this property is preserved well in the identified models. This is illustrated in Figs. 5 and 6, which contain the optimal approximations of orders 2 and 4.

We also apply the GTLS method to data w_n consisting of two noisy steps for the input and the corresponding system output. These data and the optimal approximation of order 4

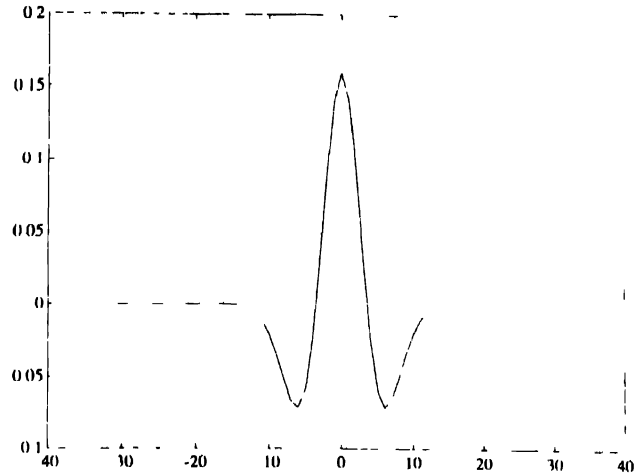


Fig. 4 The Mexican hat.

TABLE VIII
MODEL QUALITY

model		B_{gtls}	B_{bal}	B_{hank}
n=2	misfit	0.17	0.18	0.32
	error in impulse response	0.21	0.20	0.34
n=4	misfit	0.046	0.047	0.052
	error in impulse response	0.051	0.050	0.055
n=6	misfit	0.0070	0.0071	0.0078
	error in impulse response	0.0076	0.0075	0.0081

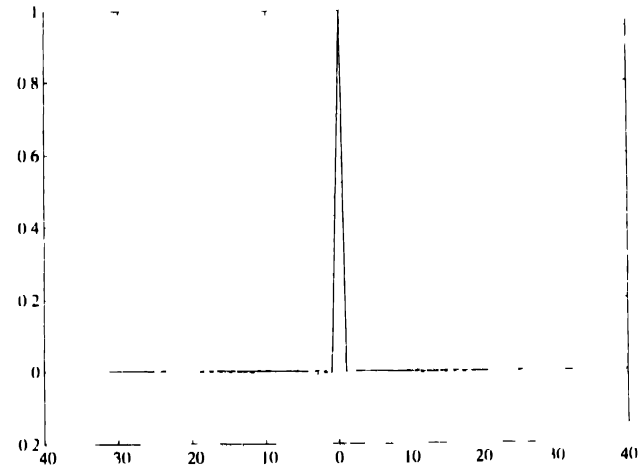


Fig. 5 The pulse w_1 (solid line), the first component of the l_2 -optimal approximation of second order (dashed line) and fourth order (dash-dotted line)

are given in Fig. 7. We should mention that the approximation error in the input is so small that it is nearly invisible in this figure.

The misfit of this model and the error in its impulse response are listed in the first column of Table IX. In view of the results for $n = 4$ in Table VIII, this shows that the identified model is a rather accurate approximation of the Mexican hat. Depending on the choice of an initial model, it typically takes a few hundred iterations to obtain convergence, and sometimes convergence to a local optimum occurred.

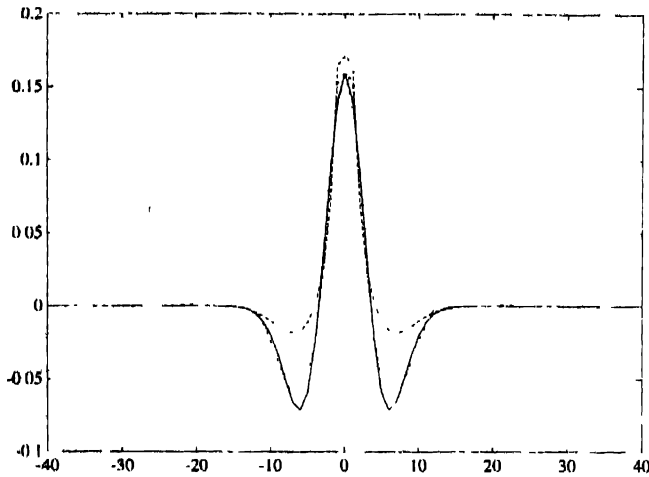


Fig. 6. The Mexican hat w_2 (solid line), the second component of the l_2 -optimal approximation of second order (dashed line) and fourth order (dash-dotted line).

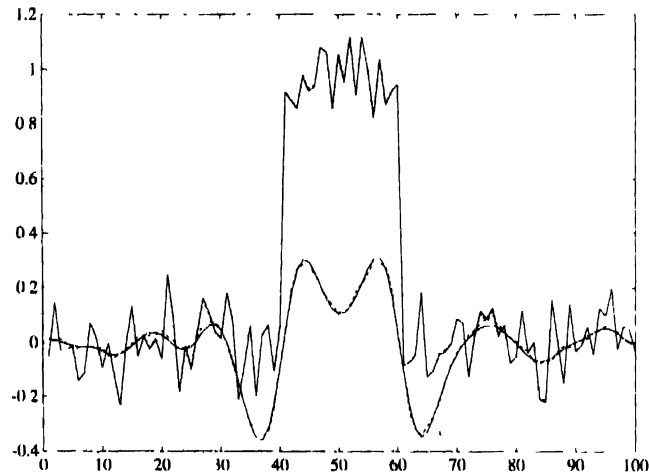


Fig. 7. Noisy step measurement w_n (solid lines) and its l_2 -optimal approximation of fourth order (dashed lines).

TABLE IX
LOCAL VS GLOBAL METHOD

method	Global TLS	Local TLS
misfit with respect to w_n	0.11	1.17
error in impulse response	0.066	0.522

Finally, as mentioned at the end of Section VI, we will once more consider the local total least squares method described in Section III. The results for the fourth order model are given in the second column in Table IX. This clearly shows that, perhaps not surprisingly, the local method gives poor results with respect to the global total least squares criterion. For example, the error in the impulse response of the local model is even larger than the Mexican hat itself, which has norm $\|w_2\| = 0.35$.

C. A System with Multiple Outputs

In this experiment we consider a system with multiple outputs, so that a single difference equation does not suffice to describe the system. For simplicity we consider a system

TABLE X
SELECTION OF RANK AND DEGREE

model	$n = 0$	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$m = 1$	3.33	1.45	0.93	0.89	0.88
$m = 2$	1.30	0.61	0.56	0.52	0.46

with one input and two outputs. The data are generated as $w = w' + v$, where $w' \in \mathcal{B}^{3,1,2}$ satisfies the equations

$$\begin{aligned} w'_2(t) &= w'_2(t-1) + w'_1(t) \\ w'_3(t) &= w'_2(t) + w'_1(t-1). \end{aligned} \quad (25)$$

For w_1 we take white noise with unit variance, and for v a three-dimensional white noise process with independent components and variance 0.01. The observation interval has length 50. Outside this interval we define $w(t) = 0$. The GTLS model of rank one and degree two is

$$\begin{aligned} w_2(t) &= 0.95w_2(t-1) - 0.04w_2(t-2) + 1.08w_1(t) \\ &\quad + 0.07w_1(t-1) + 0.15w_1(t-2) \\ w_3(t) &= 0.95w_3(t-1) - 0.04w_3(t-2) + 1.12w_1(t) \\ &\quad + 1.08w_1(t-1) - 0.89w_1(t-2). \end{aligned} \quad (26)$$

Transforming model equation (25) to the form

$$\begin{aligned} w'_2(t) - w'_2(t-1) &= w'_1(t) \\ w'_3(t) - w'_3(t-1) &= w'_1(t) + w'_1(t-1) - w'_1(t-2) \end{aligned} \quad (27)$$

shows that the original model equations are estimated rather accurately. We compare the misfits of GTLS models of various complexity in Table X.

For rank one, the misfit hardly decreases for orders above two. This could be expected, as the regular part of the data belongs to a system of order two. For rank two the results suggest to take the order one. Comparing the complexities $(m, n) = (1, 2)$ and $(2, 1)$, the first one of course leads to a larger misfit, as it imposes more restrictions. The misfit is still relatively small, however, when compared to the norm of the data $\|w\| = 6.10$.

X. CONCLUSION

In this paper we investigated the modeling of vector time series by means of difference equations, using the global total least squares criterion. Distinctive features of our approach are that no decomposition into inputs and outputs is required and that the criterion measures the global misfit in a nonparametric way. The misfit of a given system is evaluated by a dynamic projection algorithm formulated in terms of isometric state representations. We developed an iterative algorithm for constructing optimal models and gave a characterization of stationary points of the GTLS criterion. The method was illustrated by some simulation experiments.

The results of this paper can be extended in several directions, e.g., time series on a finite time interval, time varying systems, and time-varying norms for the misfit. Further research will be concerned with statistical properties and the development of faster algorithms and recursive methods.

APPENDIX A l₂-SYSTEMS

By l_2^q we denote the set of q -dimensional square summable time series over time axis \mathbf{Z} , i.e., $l_2^q := \{w: \mathbf{Z} \rightarrow \mathbf{R}^q; \sum_{t=-\infty}^{\infty} w(t)^T w(t) < \infty\}$. We define l_2 -systems as follows. By σ we denote the time shift defined by $(\sigma w)(t) := w(t+1)$, $t \in \mathbf{Z}$.

Definition A.1 (l_2 -systems): An l_2 -system is a subset \mathcal{B} of l_2^q that is:

- 1) linear, i.e., for $w, w' \in \mathcal{B}$ and $\alpha, \beta \in \mathbf{R}$ there holds $\alpha w + \beta w' \in \mathcal{B}$;
- 2) shift-invariant, i.e., $\sigma \mathcal{B} = \mathcal{B}$;
- 3) complete, i.e., if $w \in l_2^q$ satisfies $w|_T \in \mathcal{B}|_T$ for all finite $T \subset \mathbf{Z}$, then $w \in \mathcal{B}$.

This class of systems can be represented by linear, time-invariant difference equations of finite lag. Let $\mathbf{R}^{p \times q}[s]$ denote the set of $p \times q$ matrices with polynomial entries. A set of p difference equations corresponds to a polynomial matrix in the shift σ , i.e., $R(\sigma) \in \mathbf{R}^{p \times q}[\sigma]$.

Proposition A.2: A set $\mathcal{B} \in l_2^q$ is an l_2 -system if and only if $\mathcal{B} = \{w \in l_2^q; R(\sigma)w = 0\}$ for some $R \in \mathbf{R}^{p \times q}[s]$ and $p \in \mathbf{N}$.

Proof: The if-part is trivial, and the other part is proved by construction; see [10, Theorem 5]. \square

The complexity of a system is measured in terms of its dimension. Considered as a linear space, every nonzero l_2 -system has infinite dimension. Therefore we consider its dimension on finite time intervals.

Definition A.3: For a given system \mathcal{B} let $\mathcal{B}^0 := \{w \in \mathcal{B}; w(t) = 0 \text{ for } t < 0\}$, and let \mathcal{B}_0^0 be the restriction of \mathcal{B}^0 to time $t = 0$.

- 1) The rank of \mathcal{B} is defined as $m(\mathcal{B}) := \dim(\mathcal{B}_0^0)$.
- 2) The degree of \mathcal{B} is defined as $n(\mathcal{B}) := \dim(\mathcal{B}|_{[0, \infty)} \bmod \mathcal{B}_0^0|_{[0, \infty)})$.

The rank and degree determine the dimension of an l_2 -systems on finite intervals. To be specific, if \mathcal{B} is a system with rank m and degree n , then $\dim(\mathcal{B}|_T) = mN + n$ for intervals $T \subset \mathbf{Z}$ of length $N \geq n$. The rank and degree have the following interpretation. The rank is the number of degrees of freedom for a system at each time instant, given the past. This is equal to the number of inputs in the system. The degree measures the remaining freedom due to initial conditions. This is equal to the number of states. This is made precise in the following result.

Proposition A.4: The rank and degree of a system equal, respectively, the number of auxiliary inputs and the number of states in a minimal state representation.

Proof: For a proof we refer to [10, Theorem 9]. \square

APPENDIX B PROOFS

Proof of Proposition 4.3:

Necessity of 1: Let $\mathcal{R} := \text{im}[B \ AB \ \dots \ A^{n-1}B]$, then $A\mathcal{R} \subset \mathcal{R}$ and $\text{im } B \subset \mathcal{R}$. If (A, B) is not controllable there is a choice of basis for the state space such that with respect to this basis A and B take the form $\begin{bmatrix} A_1 & A_2 \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} B_1 \\ 0 \end{bmatrix}$.

So in a corresponding partition $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ of the state space, it holds that $\sigma x_2 = A_2 x_2$. As this admits only the zero solution in l_2 , x_2 can be removed.

Necessity of 2: Suppose there exists an f such that $(A + BF, C + DF)$ is not observable. Then the unobservable components of the state can be removed, as in the previous part.

Necessity of 3: An obvious condition is that $\begin{bmatrix} B \\ D \end{bmatrix}$ has full column rank, so it suffices to prove that $\ker D \subset \ker B$. Suppose $\ker D \setminus \ker B \neq \{0\}$, then there exists an invertible $R \in \mathbf{R}^{m \times m}$ such that $\begin{bmatrix} BR \\ DR \end{bmatrix} = \begin{bmatrix} B' & b \\ D' & 0 \end{bmatrix}$ with $0 \neq b \in \mathbf{R}^n$. Then \mathcal{B} is represented by $\sigma x = Ax + B'u' + bz$ and $w = Cx + D'u'$. As z influences w only with a delay, we remove its direct influence on σx by defining $\sigma x' := \sigma x - bz$. This gives $\sigma x' = Ax' + B'u' + Ab\sigma^{-1}z$ and $w = Cx' + D'u' + C'b\sigma^{-1}z$. Hence (A, B, C', D) is equivalent to $(A, [B' \ Ab], C', [D' \ C'b])$. From Proposition 4.2 it follows that this is equivalent to $(A + Abf, [B' \ Ab], C' + C'bf, [D' \ C'b])$, for all $f \in \mathbf{R}^{1 \times n}$. As $b \neq 0$, f can be chosen such that $I_n + bf$ is singular, by taking $f = -b^T / \|b\|^2$. It is easily verified that then $(A + Abf, C' + C'bf) = (A(I_n + bf), C'(I_n + bf))$ is not observable. From the Necessity of 2 it follows that the state dimension can be reduced.

Sufficiency of the Conditions: Let (A, B, C', D) be an SR with m auxiliary inputs and n states that satisfies Conditions 1), 2), and 3). From controllability and Proposition 4.2 it follows that without loss of generality we may assume that A is asymptotically stable. We prove that the rank of \mathcal{B} equals m , and its degree equals n . Then the result follows from Proposition A.4.

Concerning the rank, consider \mathcal{B}^0 as introduced in Definition A.3. Observability implies that trajectories $w \in \mathcal{B}^0$ have state zero at time $t = 0$. So $\mathcal{B}_0^0 = \text{im } D$, and from Condition 3) it follows that $\dim \mathcal{B}_0^0 = \text{rank } D = m$. Concerning the degree, we consider the space $\mathcal{B}|_{[0, \infty)} \bmod \mathcal{B}_0^0|_{[0, \infty)}$; cf. Definition A.3-2). This space can be parameterized by the initial state x_0 . Observability implies that this parameterization is injective, from which the result follows. \square

Proof of Proposition 5.2: Write $(A', B', C', D') := (S(A + BF)S^{-1}, SBR, (C + DF)S^{-1}, DR)$. Then

$$\begin{pmatrix} A' & B' \\ C' & D' \end{pmatrix} = \begin{pmatrix} S & 0 \\ 0 & I_m \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} S^{-1} & 0 \\ FS^{-1} & R \end{pmatrix}.$$

Equation (13) for (A', B', C', D') gives, with $K = S^T S$

$$\begin{aligned} & \begin{pmatrix} A'^T & C'^T \\ B'^T & D'^T \end{pmatrix} \begin{pmatrix} K & 0 \\ 0 & I_m \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \\ &= \begin{pmatrix} S^{-1} & 0 \\ FS^{-1} & R \end{pmatrix}^{-T} \begin{pmatrix} S^{-1} & 0 \\ FS^{-1} & R \end{pmatrix}^{-1} \\ &= \begin{pmatrix} S^T & -F^T R^{-T} \\ 0 & R^{-T} \end{pmatrix} \begin{pmatrix} S & 0 \\ -R^{-1}F & R^{-1} \end{pmatrix} \\ &= \begin{pmatrix} K + F^T(RR^T)^{-1}F & -F^T(RR^T)^{-1} \\ -(RR^T)^{-1}F & (RR^T)^{-1} \end{pmatrix}. \end{aligned}$$

Now verification of (16), (17), and (14) is straightforward. For a proof of the uniqueness of K we refer to [8]. \square

Proof of Proposition 5.3: From (13) it follows that $A^T A = I_n - C^T C$, hence $|Ax|^2 = |x|^2 - |Cx|^2 \leq |x|^2$. So A is stable. We prove that the representation is not minimal if A is not asymptotically stable. In that case A has an eigenvalue λ with $|\lambda| = 1$. Let x denote a corresponding eigenvector, and x^* its complex conjugate. Then $|Cx|^2 = x^* C^T C x = x^* x - x^* A^T A x = |x|^2 - |Ax|^2 = 0$. This implies that $CA^k x = 0$, $k \geq 0$, so that (A, C) is not observable and hence not minimal.

From the proof of Proposition 5.2 it follows that (15)–(17) are necessary conditions. Equation (15) determines S modulo a left unitary factor, corresponding to U . Equation (16) determines R modulo a right unitary factor, corresponding to V . \square

Proof of Theorem 6.2: From Definition 5.1 it follows that (A, \bar{B}, C, \bar{D}) as defined in part 4) is an ISR. As A is asymptotically stable [see Proposition 5.3-1)] the corresponding image representation $\bar{G}: l_2^{q-m} \rightarrow l_2^q$ is well defined. We first prove that

$$GG^* + \bar{G}\bar{G}^* = I_q \quad (28)$$

Define $\bar{B} = [B \ \bar{B}]$ and $\bar{D} = [D \ \bar{D}]$. Then (A, \bar{B}, C, \bar{D}) is also an isometric state representation. Let $\bar{G}(\sigma): l_2^q \rightarrow l_2^q$ denote the corresponding image representation. Then $\bar{G} = [G \ \bar{G}]$, so $GG^* + \bar{G}\bar{G}^* = \bar{G}\bar{G}^*$. It remains to prove that $\bar{G}\bar{G}^* = I_q$, or equivalently, that $\bar{G}^* = \mathcal{R}_q \bar{G}^T \mathcal{R}_q$ is isometric. The operator \mathcal{R}_q is clearly isometric, and because \bar{G}^T is the image representation induced by the ISR $(A^T, C^T, \bar{B}^T, \bar{D}^T)$, it is also isometric. This proves (28).

- 1) From Theorem 6.1 it follows that $\bar{G}\bar{G}^* w$ is the optimal approximation of w in $\mathcal{B}(A, \bar{B}, C, \bar{D})$. From (28) we obtain $\{w \in \mathcal{B}(A, \bar{B}, C, \bar{D})\} \Leftrightarrow \{\bar{G}\bar{G}^* w = w\} \Leftrightarrow \{GG^* w = 0\} \Leftrightarrow \{w \in \mathcal{B}^\perp\}$. So $\text{im } \bar{G} = \mathcal{B}^\perp$, which proves 1).
- 2) Equation (28) shows that every $w \in l_2^q$ can be decomposed into a part contained in \mathcal{B} and one in \mathcal{B}^\perp .
- 3) As (A, \bar{B}, C, \bar{D}) is a state representation of \mathcal{B}^\perp , it follows that $m(\mathcal{B}^\perp) \leq q - m$ and $n(\mathcal{B}^\perp) \leq n$. From part 2) it follows that $m(\mathcal{B}^\perp) \geq q - m$, so $m(\mathcal{B}^\perp) = q - m$. As $(\mathcal{B}^\perp)^\perp = \mathcal{B}$, there holds that $n = n((\mathcal{B}^\perp)^\perp) \leq n(\mathcal{B}^\perp)$, so that $n(\mathcal{B}^\perp) = n$. This proves 3).
- 4) That (A, \bar{B}, C, \bar{D}) is an ISR of \mathcal{B}^\perp was proved in 1), and minimality follows from 3) and Proposition A.4.
- 5) Equation (28) implies that $\hat{w} = w - \bar{w} = w - GG^* w = \bar{G}\bar{G}^* w$, and the result follows from Theorem 6.1. \square

Proof of Theorem 8.1: It suffices to prove the theorem for a minimal isometric SR. This can be seen as follows. If condition 2) holds for an arbitrary minimal SR, then it holds for all equivalent minimal SR's, as in the transformation $(S(A + BF)S^{-1}, SBR, (C + DF)S^{-1}, DR)$, the auxiliary input and state are linearly transformed to $R^{-1}(\hat{v} - F\hat{x})$ and Sx ; see the proof of Proposition 4.2.

So let (A, B, C, D) be a minimal ISR of \mathcal{B} , and let \bar{B}, \bar{D} , be defined as in Theorem 6.2-4). In the proof we will make use of the following relations, which follow from the projection

scheme in Section VI; see Fig. 2.

$$\text{for } \hat{v}: x = A^T \sigma x + C^T w, \quad \hat{v} = B^T \sigma x + D^T w \quad (29)$$

$$\text{for } \tilde{v}: x = A^T \sigma x + C^T w, \quad \tilde{v} = \bar{B}^T \sigma x + \bar{D}^T w \quad (30)$$

$$\text{for } \hat{w}: \sigma \hat{x} = A \hat{x} + B \hat{v}, \quad \hat{w} = C \hat{x} + D \hat{v} \quad (31)$$

$$\text{for } \tilde{w}: \sigma \tilde{x} = A \tilde{x} + \bar{B} \tilde{v}, \quad \tilde{w} = C \tilde{x} + \bar{D} \tilde{v}. \quad (32)$$

Further, from (13) we obtain

$$\begin{aligned} \hat{x} &= A^T \sigma \hat{x} + C^T \hat{w}, & \hat{v} &= B^T \sigma \hat{x} + D^T \hat{w}, \\ 0 &= \bar{B}^T \sigma \hat{x} + \bar{D}^T \hat{w} \end{aligned} \quad (33)$$

$$\begin{aligned} \tilde{x} &= A^T \sigma \tilde{x} + C^T \tilde{w}, & 0 &= B^T \sigma \tilde{x} + D^T \tilde{w}, \\ \tilde{v} &= \bar{B}^T \sigma \tilde{x} + \bar{D}^T \tilde{w}. \end{aligned} \quad (34)$$

We first describe those optimality conditions that can be derived straightforwardly from the model improvement constructions in Theorem 7.2.

Lemma B.1:

- 1) Construction 1 gives no improvement if and only if $\text{cov}([\hat{x}, \hat{v}], \hat{w}) = 0$.
- 2) Construction 2 gives no improvement if and only if $\text{cov}(\hat{v}, [\sigma \tilde{x}, \tilde{w}]) = 0$.
- 3) If Construction 3 gives no improvement, then $\text{cov}(\hat{v}, \tilde{v}) = 0$.

Proof of the Lemma: We use the notation of Theorem 7.2.

- 1) As $\hat{w} = C\hat{x} + D\hat{v}$, clearly $\hat{w} \in \mathcal{E}$. Construction 1 does not yield an improvement iff the projection of \hat{w} onto \mathcal{E} is zero, from which the result follows.
- 2) Construction 2 gives no improvement iff the projection of \hat{w} onto F is zero, cf. part 1). This is equivalent to $\langle \hat{w}, (C(\sigma I - A)^{-1} B' + D')v \rangle = 0$ for all B', D' . From (34) this is equivalent to $\langle \hat{w}, \sum_{k=0}^{\infty} C^k A^{k-1} B' \sigma^{-k} \hat{v} + D' \hat{v} \rangle = \langle \sum_{k=1}^{\infty} B'^T (A^T)^{k-1} C^T \sigma^k \hat{w} + D'^T w, v \rangle = \langle B'^T \sigma \tilde{x} + D'^T \tilde{w}, v \rangle = 0$, from which the result follows.
- 3) If $\text{cov}(\hat{v}, \tilde{v}) \neq 0$, then construction 3 decreases $\|\hat{v}\| = \|\tilde{v}\|$. \square

Returning to Theorem 8.1, we first prove part 2. From the lemma it follows that $\text{cov}(\hat{v}, \tilde{v}) = 0$, and $\text{cov}(\hat{v}, \hat{x}) = \text{cov}(\hat{v}, A^T \sigma \hat{x} + C^T \hat{w}) = 0$. As \bar{D} is injective, $\text{cov}(\hat{x}, \tilde{v}) = 0$ if and only if $\text{cov}(\hat{x}, \bar{D}\tilde{v}) = 0$, which is equivalent to $\text{cov}(\hat{x}, \tilde{w} - C\tilde{x}) = -\text{cov}(\hat{x}, C\tilde{x}) = 0$. So to prove 2 it remains to show that $\text{cov}(\hat{x}, \hat{x}) = 0$. As (A^T, B^T) is observable, this is equivalent to $\text{cov}(\sigma \hat{x}, B^T A^T \sigma \hat{x}) = 0$ for all $k \geq 0$. We prove this by induction. For $k = 0$, the lemma shows that $\text{cov}(\sigma \hat{x}, B^T \sigma \hat{x}) = \text{cov}(\sigma \hat{x}, -D^T \hat{w}) = 0$. Now suppose that $\text{cov}(\sigma \hat{x}, B^T A^T \sigma \hat{x}) = 0$ for $k \leq N$. Then $\text{cov}(\sigma \hat{x}, B^T A^T \sigma \hat{x}) = \text{cov}(A\hat{x} + B\hat{v}, B^T A^T \sigma \hat{x}) = \text{cov}(A\hat{x}, B^T A^T \sigma \hat{x}) = \text{cov}(A\hat{x}, B^T A^T \sigma (\hat{x} - C^T \hat{w})) = A \text{cov}(\hat{x}, B^T A^T \sigma \hat{x}) = 0$. This proves Theorem 8.1.2. Concerning the equivalence of 1, 2, and 3, the implications $3 \Rightarrow 2 \Rightarrow 1$ are trivial. Further, if 1 holds then $\text{cov}(\hat{x}, \hat{x}) = \text{cov}(\sigma \hat{x}, \sigma \hat{x}) = \text{cov}(A\hat{x} + B\hat{v}, A\hat{x} + B\hat{v}) = \text{cov}(A\hat{x}, A\hat{x}) = A \text{cov}(\hat{x}, \hat{x}) A^T$. As A is asymptotically stable, it follows that $\text{cov}(\hat{x}, \hat{x}) = 0$, so 1 implies 2. Finally, 3 is easily derived from 2 by using (31) and (32). \square

Proof of Theorem 8.2: Let w be a given observation and let \hat{w} be its optimal approximation in \mathcal{B} . Further let

(A, B, C, D) be a minimal SR of B with A asymptotically stable, and write the approximation error as $\tilde{w} := w - \hat{w} = w - (C(\sigma I - A)^{-1}B + D)\hat{v}$. We have to prove that B satisfies the optimality conditions if and only if the derivative of $\|\tilde{w}\|$ with respect to the parameters in A, B, C, D , and \hat{v} is zero. First we analyze the tangent space of \tilde{w} with respect to these parameters. Let \mathcal{E} and \mathcal{F} be defined as in Theorem 7.2, and let $\mathcal{G} := \{\bar{w} \in l_2^q; \exists H \in \mathbb{R}^{n \times n} \text{ such that } \bar{w} = C(\sigma I - A)^{-1}H\hat{x}\}$ with \hat{x} the state corresponding to \hat{w} .

Lemma B.2: The tangent space \mathcal{T} of $\tilde{w} = w - (C(\sigma I - A)^{-1}B + D)\hat{v}$ is given by $\mathcal{T} = \mathcal{B} + \mathcal{E} + \mathcal{F} + \mathcal{G}$.

Proof of the Lemma: The tangent space is defined as the smallest closed subspace of l_2 containing all partial derivatives. Note that \tilde{w} is linear in \hat{v} , in B , in C and in D . A change of \hat{v} corresponds to adding $\bar{w} \in \mathcal{B}$ to \tilde{w} , a change of C and D to adding $\bar{w} \in \mathcal{E}$ to \tilde{w} and a change of B and D to adding $\bar{w} \in \mathcal{F}$ to \tilde{w} . It remains to prove that the derivatives of \tilde{w} with respect to the parameters in A span the space \mathcal{G} . For $H \in \mathbb{R}^{n \times n}$ let x' be defined by $\sigma x' = (A - H)x' + B\hat{v}$ and let $w' := Cx' + D\hat{v}$. The corresponding error is $\tilde{w}' := w - w'$, so that the change in \tilde{w} is given by $\bar{w} = \tilde{w}' - \tilde{w} = \hat{w} - w' = C(\hat{x} - x') = C\bar{x}$ for $\bar{x} := \hat{x} - x'$. As $\sigma \bar{x} = A\hat{x} - (A - H)x' = A\bar{x} + Hx' = A\bar{x} + H\hat{x} - H\bar{x}$, ignoring the second-order term $H\bar{x}$ for small H gives the result. \square

We will next prove the theorem by showing that the optimality conditions and the stationarity condition are both equivalent to $\tilde{w} \perp \mathcal{T}$.

Stationarity is equivalent to the condition that $\lim_{\delta \rightarrow 0} \delta^{-1} \{\|\tilde{w} + \delta \bar{w}\| - \|\tilde{w}\|\} = 0$ for all $\bar{w} \in \mathcal{T}$. It is easily verified that this limit equals $\langle \tilde{w}, \bar{w} \rangle / \|\bar{w}\|$, so stationarity is equivalent to $\tilde{w} \perp \mathcal{T}$.

Finally we show that $\tilde{w} \perp \mathcal{T}$ is equivalent to the optimality conditions. First, suppose that the optimality conditions hold. As \tilde{w} is an optimal approximation within \mathcal{B} this holds that $\tilde{w} \perp \mathcal{B}$. Further, Theorem 8.1-3) states that $\text{cov}([\hat{v}, \hat{x}], \tilde{w}) = 0$, so that $\tilde{w} \perp \mathcal{E}$, and $\text{cov}(\hat{v}, [\sigma \bar{x}, \tilde{w}]) = 0$, so that the proof of Lemma B.1-2) shows that $\tilde{w} \perp \mathcal{F}$. Finally, for $\bar{w} \in \mathcal{G}$ given by $\bar{w} = \sum_{k=1}^{\infty} CA^{k-1}H\sigma^{-k}\hat{x}$ we obtain by using (34) that $\langle \bar{w}, \tilde{w} \rangle = \langle \sum_{k=1}^{\infty} CA^{k-1}H\sigma^{-k}\hat{x}, \tilde{w} \rangle = \langle \hat{x}, \sum_{k=1}^{\infty} H^T A^{T(k-1)} C^T \sigma^k \tilde{w} \rangle = \langle \hat{x}, H^T \sigma \tilde{x} \rangle = 0$, so $\tilde{w} \perp \mathcal{G}$. From Lemma B.2 it follows that $\tilde{w} \perp \mathcal{T}$.

Second, supposing that $\tilde{w} \perp \mathcal{T}$ we prove the optimality conditions. The fact that $\tilde{w} \perp \mathcal{E} + \mathcal{F}$ implies the conditions in Lemma B.1-2) and B.1-2), cf. the proof of that lemma. The condition in Lemma B.1-3) follows from that in B.1-2) by using (34). Further, the optimality conditions were derived from these conditions in the proof of Theorem 8.1. \square

Proof of Proposition 8.4: First we prove that \mathcal{Z} is a linear space. Observe that for every $z \in \mathcal{B}^\perp$ the optimal approximation of $\hat{w} + z$ within \mathcal{B} is given by \hat{w} . Let (\hat{v}, \hat{x}) be the auxiliary input and state for \hat{w} in a minimal SR of B , and let (\hat{v}_z, \hat{x}_z) be defined analogously for z in \mathcal{B}^\perp . As (\hat{v}, \hat{x}) is fixed, the condition in Theorem 8.1-2) for stationarity of B with respect to $\hat{w} + z$ consists of linear restrictions on (\hat{v}_z, \hat{x}_z) . As \hat{x}_z is a linear function of \hat{v}_z , these conditions can be expressed as linear restrictions on \hat{v}_z alone. This shows that B is a stationary point for $\hat{w} + z$ if and only if the

auxiliary input \hat{v}_z is restricted to a linear subspace of l_2 , from which the linearity of \mathcal{Z} follows. Next we show that the minimum in (22) is achieved by taking $w_0 = w - \bar{w}_0 = \hat{w} + \tilde{w} - (\tilde{w} - \tilde{w}') = \hat{w} + \tilde{w}'$ and $w' \in \mathcal{Z}$. By definition of \mathcal{Z} it follows that B is stationary for $w - \bar{w}_0$, by definition of \mathcal{Z} . Further, $\bar{w} \in \mathcal{B}^\perp$ is such that B is stationary for $w - \bar{w} = w - \bar{w}_0$ if and only if $z := \tilde{w} - \bar{w} \in \mathcal{Z}$. Now the norm of $w - w_0$ is minimized by taking $z = \tilde{w}'$, the orthogonal projection on \tilde{w} on \mathcal{Z} , hence $\bar{w}_0 = \tilde{w} - \tilde{w}'$. \square

REFERENCES

- [1] P. E. Caines, *Linear Stochastic Systems*. New York: Wiley, 1988.
- [2] K. Glover, "All optimal Hankel norm approximations of linear multivariable systems and their L^∞ error bounds," *Int. J. Contr.*, vol. 49, no. 6, pp. 1115-1193, 1984.
- [3] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1983.
- [4] E. J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*. New York: Wiley, 1988.
- [5] C. Heij, *Deterministic Identification of Dynamical Systems* (Lecture Notes in Control and Information Sciences), vol. 127. Berlin: Springer-Verlag, 1989.
- [6] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
- [7] B. C. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. Automat. Contr.*, vol. AC-26, no. 1, pp. 17-32, 1981.
- [8] H. J. Payne and L. M. Silverman, "On the discrete time algebraic Riccati equation," *IEEE Trans. Automat. Contr.*, vol. AC-18, no. 3, pp. 226-234, 1973.
- [9] J. T. Spanos, M. H. Milman, and D. L. Mingori, "A new algorithm for L_2 optimal model reduction," *Automatica*, vol. 28, no. 5, pp. 897-909, 1992.
- [10] J. C. Willems, "From time series to linear system, part I: Finite dimensional linear time invariant systems, Part II: Exact modelling; part III: Approximate modelling," *Automatica*, vol. 22/23, nos. 5, 6, 1, pp. 561-580, 675-694, 87-115, 1986, 1987.
- [11] ———, "Paradigms and puzzles in the theory of dynamical systems," *IEEE Trans. Automat. Contr.*, vol. 36, no. 3, pp. 259-294, 1991.
- [12] J. C. Willems and C. Heij, "Scattering theory and approximation of linear systems," *Modeling, Identification and Robust Control*, C. Byrnes and A. Lindquist, Eds. Amsterdam: North-Holland, 1986, pp. 397-411.



Christiaan Heij (M'88) was born in Arnhem, The Netherlands, in 1957. He studied econometrics, mathematics, and philosophy and received the Ph.D. degree from the University of Groningen, in 1988, for a dissertation on identification of linear systems.

Since 1989 Dr. Heij has been an Assistant Professor at the Econometric Institute of the Erasmus University in Rotterdam. His research interests are in the area of modeling and identification, in particular the theory and applications of linear systems and econometrics.



Berend Roorda was born in Assen, The Netherlands, in 1965. He obtained the M.Sc. degree from the Department of Mathematics of the University of Groningen, The Netherlands, in 1989. He is currently finishing his Ph.D. dissertation on approximate modeling by the global total least squares method.

He was researcher at the PTT Research Telecommunications Laboratory in Groningen in 1989. Since 1990 he has been working as a Research Assistant at the Tinbergen Institute and the Econometric Institute of the Erasmus University Rotterdam.

Control of Chained Systems

Application to Path Following and Time-Varying Point-Stabilization of Mobile Robots

Claude Samson

Abstract—Chain form systems have recently been introduced to model the kinematics of a class of nonholonomic mechanical systems. The first part of the study is centered on control design and analysis for nonlinear systems which can be converted to the chain form. Solutions to various control problems (open-loop steering, partial or complete state feedback stabilization) are either recalled, generalized, or developed. In particular, globally stabilizing time-varying feedbacks are derived, and a discussion of their convergence properties is provided. Application to the control of nonholonomic wheeled mobile robots is described in the second part of the study by considering the case of a car-pulling trailers.

1. INTRODUCTION

FROM a theorem due to Brockett [5], it is known that nonholonomic wheeled mobile robots with restricted mobility (such as unicycle-type and car-like vehicles) cannot be stabilized to a desired configuration (or posture) via differentiable, or even continuous, pure-state feedback [3], [24], [1]. Nonsmooth feedback has been proposed as an alternative solution (see [4], [6], [32], for example). Another alternative, first pointed out by the author in [25], consists of using smooth time-varying feedbacks, i.e., feedbacks which explicitly depend on the time variable. Such feedbacks had previously been very little studied in Control Theory. The result given in [25], for a unicycle-type vehicle, has subsequently motivated research work to explore the potentialities of time-varying feedbacks [8], [9], [11], derive explicit design methods [19], [34], and extend their use in robotics applications [26], [29].

The possibility of modeling the kinematic equations of wheeled mobile robots by so-called canonical chain form equations (a particular class of nonlinear nilpotent systems) has been pointed out in [17] when treating the case of car-like vehicles. It was known before that the equations of unicycle-type vehicles (a simpler case) could be written in this form, but this had not been used explicitly at the control design level. More recently, it has been shown [31] that the equations of vehicles with trailers could also be locally converted into a chain form.

In [17], the authors aimed essentially at exploring methods for open-loop steering of nonholonomic systems by using sinusoidal inputs. More recently, the authors of [34] have realized that chained systems could also be put under another canonical form, called "power form," and that power form systems belonged to the class of systems considered in [19] for which explicit smooth time-varying stabilizing feedbacks can be derived. The method proposed in [34], for deriving such controls, is applied to the problem of locally stabilizing a car-like system to a desired posture. A global solution to this problem had previously been given by the author in [26], [27] by using another approach.

Since the power form is mathematically equivalent to the chain form, one may question the interest of using one form rather than the other one when dealing with practical applications and mobile robot control in particular. To the author's knowledge, no single and definitive answer can be brought to this question because the two forms present complementary advantages and drawbacks. For example, it seems that model equations of nonholonomic vehicles using physical coordinates, such as Cartesian coordinates and angles between articulated bodies, are naturally closer to the chain form than to the power form. In this respect, the logic of using the chain form when addressing the problem of path following should be evident from the present paper. Concerning the theoretically more difficult problem of point stabilization, the answer is not as clear. As shown in this paper, a rather simple solution to smooth feedback stabilization can be obtained by complementing path following solutions, thus making the chain form intuitively attractive in this case. Smooth time-varying stabilization, however, does not seem to be compatible with fast convergence [11]. For this reason, recent studies on point stabilization have focused on the possibility of achieving faster (exponential) convergence by using nonsmooth feedback: M'Closkey and Murray have been working with the power form (see [14] for example), with the probable reason that their approach based on the use of homogeneous system coordinates does not apply well to the chain form, while Sordalen has proposed an original solution, mixing open-loop and feedback control strategies, by using the chain form [32]. The question is therefore far from being settled at the time being and will certainly motivate future developments.

This paper is organized as follows. Section II focuses on the control of chain form systems. After pointing out some

Manuscript received July 22, 1993; revised January 27, 1994 and April 10, 1994. Recommended by Associate Editor, A. M. Bloch. This work was supported by the European ESPRIT Program, PROMotion Project 6546.

The author is with INRIA, 2004 Route de Lucioles, 06902 Sophia-Antipolis, France.

IEEE Log Number 9406099.

facts about these systems and recalling some results about the open-loop steering problem, it is shown that chain form systems can themselves be converted into a slightly different form, named here skew-symmetric chain form, particularly well adapted to subsequent Lyapunov design and analysis of globally stabilizing time-varying feedbacks. Whenever this is possible, as it is the case here, finding adequate Lyapunov functions at an early stage of the analysis presents some advantages in terms of simplicity of the control design and stability proofs. A comparison with Center Manifold techniques, as used in [34] for example, is in this respect illustrative. In the process of deriving smooth feedback control laws for chained systems, useful connections with more classical linear control techniques are carried out. Several solutions to the point-stabilization problem are then tentatively compared by analyzing the type of stability associated with each of them. This motivates a short discussion about the practical relevance of the notion of asymptotical rate of convergence.

In Section III, the results of Section II are applied to a car pulling n -trailers, seen as an extension of the unicycle and car cases. The same approach as in [29], in which stabilization to a desired configuration is treated as an extension of the path following problem, is considered. This approach involves a specific parameterization of the vehicle's posture which facilitates the decoupling of the path following problem from translational velocity control. The interest of this parameterization for solving path planning issues is also pointed out in connection with the approach developed in [21]. Finally, a modification in the modeling of the system's equations is proposed so as to broaden the control stability domain.

II. CONTROL OF CHAIN FORM SYSTEMS

A. About Chained Systems

Let us consider a chain form system which may be written as

$$\begin{aligned}\dot{x}_1 &= u_1 \\ \dot{x}_2 &= u_1 x_3 \\ \dot{x}_3 &= u_1 x_4 \\ &\vdots \\ \dot{x}_{n-1} &= u_1 x_n \\ \dot{x}_n &= u_2\end{aligned}\quad (1)$$

or, equivalently

$$\dot{X} = h_1(X)u_1 + h_2(X)u_2 \quad h_1(X) = \begin{bmatrix} 1 \\ x_3 \\ x_4 \\ \vdots \\ x_n \\ 0 \end{bmatrix} \quad h_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad (2)$$

With respect to the notations used in [17], the components of the state vector have just been ordered differently.

In the second part of the study, where system (1) will be used to model the kinematic equations of wheeled mobile robots, x_1 will represent the distance covered by the vehicle along a path to be followed, and x_2 will represent the lateral distance between the vehicle and the path. Path following will therefore mainly consist in regulating x_2 to zero, independently of the values taken by x_1 (and thus u_1) while stabilization to a desired configuration will further involve the regulation of x_1 to zero by utilizing also the input u_1 . The possibility of directly relating the first state components of the chained form to physical euclidean-type coordinates justifies the preference given here to this form over the theoretically equivalent power form introduced in [34].

It is worth noting that a chained system like (1), although it is nonlinear, has a strong underlying linear structure. This clearly appears when u_1 is taken as a function of time and no longer as a control variable. In this case, the system becomes a single-input time-varying linear system which may be written as

$$\begin{aligned}\dot{x}_1 &= 0 \\ \dot{X}_2 &= \begin{bmatrix} 0 & u_1(t) & 0 & \cdots & 0 \\ 0 & 0 & u_1(t) & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & \cdots & u_1(t) & 0 \\ 0 & \cdots & \cdots & 0 & u_1(t) \\ 0 & 0 & \cdots & \cdots & 0 \end{bmatrix} X_2 + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix} u_2\end{aligned}\quad (3)$$

with

$$\dot{x}_1 = x_1 - \int_0^t u_1(\tau) d\tau \quad \text{and} \quad X_2^T = [x_2, x_3, \dots, x_n].$$

Putting a two-input nonlinear system in the chain form, when it is possible, is thus equivalent to linearizing this system with respect to one of its inputs. Since the chained system is controllable, controllability of the original system is a necessary condition. A necessary and sufficient condition is given in [18].

When the input u_1 is taken as a function of time, the system is clearly no longer controllable due to the first equation. Under certain conditions upon the choice of $u_1(t)$, however, the second part of the system involving X_2 remains controllable.

This type of property is very useful. It is used further in the study for the derivation of smooth time-varying feedbacks which asymptotically stabilize the point $X = 0$. It can also be utilized to solve the open-loop steering problem, i.e., the problem of determining open-loop control inputs that steer the system to a desired configuration X_{desired} , chosen equal to zero without loss of generality. The method basically consists of two steps: i) choose an integrable function $u_1(t)$ which ensures controllability of the second part of the system, and determine a control $u_2(t)$ which drives $X_2(t)$ to zero in finite time (usually done by integrating the system's equations on some time interval and solving a set of algebraic equations), and ii) once X_2 is at zero, keep u_2 equal to zero so as to leave X_2 unchanged, and determine $u_1(t)$ so as to drive $x_1(t)$ to zero in finite time.

This method has been used in [16], with $u_1(t)$ and $u_2(t)$ chosen as piecewise constant inputs. In this case, the first step corresponds to discretizing the system's equations with u_1 being kept constant (and different from zero) over $n-1$ sampling time intervals Δ and applying a dead-beat control strategy (the poles of the controlled discretized system are set equal to zero) to determine the values of u_2 on the time intervals $[k\Delta, (k+1)\Delta]$ ($0 \leq k \leq n-2$). At time $t = (n-1)\Delta$, X_2 has reached zero and u_1 may then be chosen equal to $-x_1((n-1)\Delta)/\Delta$ so as to have x_1 equal to zero at time $t = n\Delta$. Note that by working more on the choice of u_1 , feedback versions of this technique can be obtained. It may also be shown, as a complement to [16], that multiplying the piecewise constant inputs by $(1 - \cos(\omega t))$, with $\omega = (2\pi/\Delta)$, does not change the values of X at the sampling instants. In this way piecewise constant inputs are transformed into time-continuous inputs that achieve the same result.

The solution proposed in [17], with $u_1(t)$ and $u_2(t)$ being composed of sinusoids at integrally related frequencies, may also be viewed as a variant of this method. Obviously, the same method applies to other inputs. A more geometrical method for open-loop steering of nonholonomic vehicles will also be pointed out further in Section III-B.

When u_1 is constant and different from zero, the above system becomes time-invariant and the second part of the system is clearly controllable. By applying classical linear control techniques, it is then possible to derive linear feedbacks $u_2(X_2)$ which stabilize the origin $X_2 = 0$ exponentially.

In fact, even if $u_1(t)$ is not constant but only piecewise continuous, bounded, and strictly positive (or negative), it is quite simple to derive stabilizing feedbacks $u_2(X_2)$ for the second part of the system. Indeed, since $x_1(t)$ varies monotonically with time, differentiation with respect to time can be replaced by differentiation with respect to x_1 . From now on we will refer to this change of variable as the u_1 -time-scaling procedure. Then, the second part of the system may equivalently be written

$$\begin{aligned} \dot{x}_2^{(1)} &= \text{sign}(u_1)x_3 \\ \dot{x}_3^{(1)} &= \text{sign}(u_1)x_4 \\ &\vdots \\ \dot{x}_{n-1}^{(1)} &= \text{sign}(u_1)x_n \\ \dot{x}_n^{(1)} &= \text{sign}(u_1)v_2 \end{aligned} \quad (4)$$

with

$$x_i^{(1)} = \text{sign}(u_1) \frac{\partial^i x_i}{\partial x_1^i} \quad \text{and} \quad v_2 = u_2/u_1(t).$$

This is the equation of a linear invariant system, an equivalent input-output representation of which is

$$x_2^{(n-1)} = \text{sign}(u_1)^{n-1} v_2. \quad (5)$$

One falls upon a controllable invariant linear system which admits exponentially stabilizing linear feedbacks in the form

$$v_2(X_2) = -\text{sign}(u_1)^{n-1} \sum_{i=1}^{n-1} g_i x_2^{(i-1)} \quad (g_i > 0, \forall i) \quad (6)$$

the control gains g_i being chosen so as to satisfy the classical Routh-Hurwitz stability criterion (the positivity of g_i is necessary but not sufficient).

Hence, the time-varying control

$$u_2(X_2, t) = u_1(t)v_2(X_2) \quad (7)$$

globally asymptotically stabilizes the origin $X_2 = 0$ in this case. Moreover, the trajectories followed by the system's solutions are invariant with respect to variations of $u_1(t)$.

This "feedback linearization" technique, associated with u_1 -time-scaling, has in fact been used by other authors working on mobile robot control. For example, Sampei *et al.* [22] have applied it to the problem of following a straight line in the case of a car pulling a single trailer. Their solution differs, however, from the one given further in the article in that they took the car's steering wheel angle as a control, instead of the angle's velocity.

In the earlier work of Dickmanns and Zapp [10], on vision-based roadline following, u_1 -time-scaling is also implicitly used together with tangent linearization of the system's equations, instead of exact feedback linearization. In their work, u_1 has the physical meaning of the car's translational velocity.

Extension of the path following problem to the point-stabilization problem to achieve smooth time-varying feedback stabilization of a unicycle-type vehicle to a given posture, based on u_1 -time-scaling, has been first proposed in [29]. The present study may be seen as a generalization of the results described in this paper.

B. Skew-Symmetric Chain Form and Lyapunov Control Design

We show next, by introducing the skew-symmetric chain form evoked before, and via a Lyapunov-like analysis, that control (7) globally stabilizes the origin $X_2 = 0$ for the second part of the chained system, provided that $|u_1(t)|$ and $|\dot{u}_1(t)|$ are bounded, and $u_1(t)$ does not asymptotically tend to zero. An important difference with the result stated previously is that $u_1(t)$ is now allowed to pass through zero.

From there it will be simple to complement the analysis and derive smooth time-varying feedbacks which globally stabilize the origin $X = 0$ of the complete system.

To this purpose, let us consider the following change of coordinates $\phi_1: X \mapsto Z$ in R^n

$$z_1 = x_1$$

$$z_2 = x_2$$

$$z_3 = x_3$$

$$z_{j+3} = k_j z_{j+1} + L_{h,j} z_{j+2} \quad 1 \leq j \leq n-3 \quad (8)$$

where

- k_j ($1 \leq j \leq n-3$) is a real positive number;
- $L_{h_1} z_j = \frac{\partial z_j}{\partial X} h_1(X)$: the Lie derivative of z_j along h_1 ;
- $L_{h_1}^k = L_{h_1}^{k-1} L_{h_1}$: the Lie differentiation operator of order k along h_1 .

One easily verifies that the Jacobian matrix $(\partial \phi_1 / \partial X)$ is a constant lower triangular matrix with ones on the diagonal. It is therefore a regular linear change of coordinates in R^n .

Moreover: $L_{h_2} z_i = 0$ ($1 \leq i \leq n-1$), and $L_{h_2} z_n = 1$.

Taking the time derivative of z_{j+3} and using 2

$$\begin{aligned} \dot{z}_{j+3} &= \frac{\partial z_{j+3}}{\partial X} \dot{X} \\ &= (L_{h_1} z_{j+3})u_1 + (L_{h_2} z_{j+3})u_2. \end{aligned} \quad (9)$$

Also, from 8

$$L_{h_1} z_{j+3} = -k_{j+1} z_{j+2} + z_{j+4}. \quad (10)$$

Hence

$$\dot{z}_{j+3} = -k_{j+1} u_1 z_{j+2} + u_1 z_{j+4} \quad (0 \leq j \leq n-4) \quad (11)$$

and

$$\dot{z}_n = L_{h_1} z_n u_1 + u_2. \quad (12)$$

The original chained system has thus been converted to the following skew-symmetric chained system

$$\begin{aligned} \dot{z}_1 &= u_1 \\ \dot{z}_2 &= u_1 z_3 \\ \dot{z}_3 &= -k_1 u_1 z_2 + u_1 z_4 \\ &\vdots \\ \dot{z}_{j+3} &= -k_{j+1} u_1 z_{j+2} + u_1 z_{j+4} \quad (0 \leq j \leq n-4) \\ \dot{z}_n &= -k_{n-2} u_1 z_{n-1} + u_2 \end{aligned} \quad (13)$$

with

$$u_2 = (k_{n-2} z_{n-1} + L_{h_1} z_n) u_1 + u_2. \quad (14)$$

The interest of this form is that it naturally lends itself to Lyapunov control design and analysis, as illustrated by the following proposition.

Proposition 2.1: Assume that $|u_1(t)|$ and $|\dot{u}_1(t)|$ are bounded, and consider the control

$$w_2 = -k_{w_2}(u_1) z_n \quad (15)$$

where $k_{w_2}(\cdot)$ is a continuous application strictly positive on $R - \{0\}$. If this control is applied to system (13), then the positive function

$$\begin{aligned} V(Z_2) &= 1/2(z_2^2 + (1/k_1)z_3^2 + (1/k_1 k_2)z_4^2 + \dots \\ &\quad + \left(1/\prod_{j=1}^{j=n-2} k_j\right) z_n^2) \end{aligned} \quad (16)$$

is nonincreasing along the closed-loop system's solutions, and asymptotically converges to some limit value V_{\lim} (which *a priori* depends on the initial conditions).

Moreover $u_1(t)V(Z_2(t))$ asymptotically tends to zero.

Therefore, if $u_1(t)$ does not asymptotically tend to zero, $V_{\lim} = 0$ and the manifold $Z_2 = 0$ is globally asymptotically stable.

The proof of this proposition, as of subsequent propositions, uses an extended version of Barbalat's Lemma stating that if a given differentiable function $f(x)$ from R^+ to R converges to some limit value when x tends to infinity, and if the derivative $(df/dx)(x)$ of this function is the sum of two terms, one being uniformly continuous and the other one tending to zero when x tends to infinity, then $(df/dx)(x)$ tends to zero when x tends to infinity.

Proof of Proposition 2.1: Taking the time derivative of V and using the system's $(n-1)$ last equations, one obtains

$$\dot{V} = \left(1/\prod_{j=1}^{j=n-2} k_j\right) z_n w_2. \quad (17)$$

Thus, if control (15) is used

$$\dot{V} = -\left(k_{w_2}(u_1)/\prod_{j=1}^{j=n-2} k_j\right) z_n^2 \quad (\leq 0). \quad (18)$$

The considered Lyapunov-like function is thus non-increasing.

This in turn implies that $\|Z_2(t)\|$ is bounded, uniformly with respect to the initial conditions. Existence and uniqueness of the system's solutions over R^+ also follows.

Now, since V is nonincreasing, $V(t)$ converges to some limit value $V_{\lim} (\geq 0)$. Since $k_{w_2}(\cdot)$ is continuous and since $|u_1(t)|$ and $|\dot{u}_1(t)|$ are bounded, $k_{w_2}(u_1(t))$ is uniformly continuous. Hence, the right-hand side member of equality (18) is uniformly continuous along any system's solution, and, by application of Barbalat's lemma, $V(t)$ tends to zero. Therefore, $k_{w_2}(u_1(t))z_n(t)$ tends to zero. This in turn implies, using the properties of the function $k_{w_2}(\cdot)$ and the boundedness of $|u_1(t)|$ and $|z_n(t)|$, that $u_1(t)z_n(t)$ tends to zero.

From now on, the time index will often be omitted to simplify the notations. Taking the time derivative of $u_1^2 z_n$ and using the convergence of $u_1 z_n$ to zero, gives

$$\frac{d}{dt}(u_1^2 z_n) = -k_{n-2} u_1^3 z_{n-1} + o(t) \quad \text{with} \quad \lim_{t \rightarrow +\infty} o(t) = 0 \quad (19)$$

$u_1^3 z_{n-1}$ is uniformly continuous along a system's solution since its time derivative is bounded. Therefore, in view of (19) and since $u_1^2 z_n$ tends to zero, $(d/dt)(u_1^2 z_n)$ also tends to zero (by application of the aforementioned extended version of Barbalat's lemma). Hence, $u_1^3 z_{n-1}$, and thus $u_1 z_{n-1}$, tend to zero.

Taking the time derivative of $u_1^2 z_j$ and repeating the above procedure iteratively, one obtains that $u_1 z_j$ tends to zero for $2 \leq j \leq n$. In view of the system's equations, we note that this in turn implies the convergence of \dot{Z}_2 to zero.

Summing up the squared values of $u_1(t)z_j(t)$, for $2 \leq j \leq n$, it appears that $u_1(t)^2 V(t)$ tends to zero. And so does $u_1(t)^2 V_{\lim}$ (from the already established convergence of $V(t)$ to V_{\lim}). \diamond

Remarks:

- One can verify that, with the particular choice $k_{w_2}(u_1) = k'_{w_2}|u_1|$, the set of controls u_2 given by (14) and (15) with $k_i > 0$ ($i = 1, \dots, n-2$) and $k'_{w_2} > 0$, coincides with the set of stable linear controls (7) previously associated with the linear invariant system (4). More precisely, there is a one-to-one correspondence between the elements of the two sets. One can thus apply classical linear control design methods to determine adequate values for the parameters k_i ($i = 1, \dots, n-2$) and k'_{w_2} and optimize the control performance near $Z_2 = 0$, as illustrated in [29]. This correspondence also underlies the connection existing between the Routh–Hurwitz criterion and the operation consisting in transforming the original chain form into a skew-symmetric chain form.
- Nonconvergence of $u_1(t)$ to zero, under the assumption that $|\dot{u}_1(t)|$ is bounded, implies that $\int_0^t |u_1(\tau)| d\tau$ tends to infinity with t . Divergence of this integral is in fact necessary to the asymptotical convergence of $\|Z_2(t)\|$ to zero, when using control (15) with $k_{w_2}(u_1) = k'_{w_2}|u_1|$. This appears clearly when interpreting this control as a stabilizing linear control for the linear invariant system (4) obtained by replacing the time variable by the aforementioned integral. This integral may still diverge, however, when $u_1(t)$ tends to zero slowly “enough” (like $t^{-\frac{1}{2}}$, for example). This indicates that $\|Z_2(t)\|$ may still converge to zero when $u_1(t)$ does.

Proposition 2.1 is not only of interest in solving the path following problem for mobile robots, it also suggests a way of determining smooth time-varying feedbacks which globally asymptotically stabilize the origin $Z = 0$ (or $X = 0$) of the whole system. In this case, u_1 is used as a control the role of which is to complement the action of the control u_2 (or w_2) in order to also obtain asymptotical convergence of z_1 (or x_1) to zero. Since chained systems like (1) cannot be asymptotically stabilized by using smooth pure state feedbacks (by application of a Brockett’s theorem [5]), smooth feedback stabilization can only be achieved by using another type of control. A time-varying control law will be considered in the present case.

Proposition 2.2: Consider the same control as in Proposition 2.1

$$w_2 = -k_{w_2}(u_1)z_n \quad (20)$$

complemented with the following time-varying control

$$u_1 = -k_{u_1}z_1 + h(Z_2, t) \quad (21)$$

where

- k_{u_1} is a positive number;
- $h(Z_2, t)$ is a function of class C^{p+1} ($p \geq 1$), uniformly bounded with respect to t , with all successive partial derivatives also uniformly bounded with respect to t , and such that $C_1: h(0, t) = 0, \forall t$ C_2 : There is a time-

diverging sequence $\{t_i\}_{i \in \mathbb{N}}$, and a positive continuous function $\alpha(\cdot)$ such that

$$\|Z_2\| \geq l > 0 \Rightarrow \sum_{j=1}^{p+1} \left(\frac{\partial^j h}{\partial t^j}(Z_2, t_i) \right)^2 \geq \alpha(l) > 0, \forall i.$$

Controls (20) and (21) globally asymptotically stabilize the origin $Z = 0$.

Proof of Proposition 2.2: It has already been shown that the positive function $V(Z_2)$ used in Proposition 2.1 is non-increasing along the closed-loop system’s solutions, implying that $\|Z_2(t)\|$ is bounded uniformly with respect to initial conditions. Note that the boundedness of $|u_1(t)|$ and $|\dot{u}_1(t)|$ is not needed to prove this fact.

The first equation of the controlled system is

$$\dot{z}_1 = -k_{u_1}z_1 + h(Z_2, t). \quad (22)$$

This is the equation of a stable linear system subjected to the bounded additive perturbation $h(Z_2(t), t)$. Therefore, $|z_1(t)|$ is also bounded uniformly with respect to the initial conditions.

Existence and uniqueness of the solutions over R^+ is thus ensured.

From the expression of u_1 , it is then found that $u_1(t)$ (taken as a function of time along a system’s solution) is bounded. And so is its first derivative [by using the regularity properties imposed upon $h(Z_2, t)$].

Proposition 2.1 thus applies. In particular, $V(Z_2(t))$ tends to some limit value $V_{\lim} (\geq 0)$, $\|\dot{Z}_2(t)\|$ tends to zero, and $Z_2(t)$ tends to zero if $u_1(t)$ does not.

We now proceed by contradiction.

Assume that $u_1(t)$ does not tend to zero. Then, $\|Z_2(t)\|$ tends to zero. By uniform continuity and since $h(0, t) = 0$ (condition C_1), $h(Z_2(t), t)$ also tends to zero. Equation (22) then becomes the equation of a stable linear system subjected to an additive perturbation which asymptotically vanishes. As a consequence, $z_1(t)$ tends to zero. From the expression of u_1 , this in turn implies that $u_1(t)$ tends to zero, yielding a contradiction.

Therefore, $u_1(t)$ must asymptotically tend to zero.

Differentiating the expression of u_1 with respect to time and using the convergence of $u_1(t)$ and $\|\dot{Z}_2(t)\|$ to zero, we get

$$\dot{u}_1(t) = \frac{\partial h}{\partial t}(Z_2(t), t) + o(t) \quad \text{with} \quad \lim_{t \rightarrow +\infty} o(t) = 0. \quad (23)$$

Since $(\partial h / \partial t)(Z_2(t), t)$ is uniformly continuous (its time derivative is bounded), $\dot{u}_1(t)$, and thus $(\partial h / \partial t)(Z_2(t), t)$, tend to zero (Barbalat’s Lemma).

By using similar arguments, one obtains that the time-derivative of $(\partial h / \partial t)(Z_2(t), t)$ and $(\partial^2 h / \partial t^2)(Z_2(t), t)$ tend to zero.

By repeating the same procedure as many times as necessary, we show that $(\partial^j h / \partial t^j)(Z_2(t), t)$ tends to zero, ($1 \leq j \leq p$). Therefore

$$\lim_{t \rightarrow \infty} \sum_{j=1}^{p+1} \left(\frac{\partial^j h}{\partial t^j}(Z_2(t), t) \right)^2 = 0. \quad (24)$$

Assume now that V_{lim} is different from zero. This implies that $\|Z_2(t)\|$ remains larger than some positive real number l (which can be calculated from V_{lim}). The previous convergence result is then not compatible with the condition C_2 imposed on the function $h(Z_2, t)$.

Therefore, V_{lim} is equal to zero, and $Z_2(t)$ asymptotically converges to zero. Then, by uniform continuity and using condition C_1 , $h(Z_2(t), t)$ tends to zero.

In view of the expression of u_1 , asymptotical convergence of $z_1(t)$ to zero readily follows. \diamond

Remark: Dependence of the function $h(Z_2, t)$ upon the last state variable z_n is not required when the function $k_{w_2}(\cdot)$ is strictly positive on R . The reason is that $z_n(t)$ unconditionally converges to zero in this case, due to the convergence of $\dot{V}(Z_2(t))$ to zero (cf. proof of Proposition 2.1).

It can be noted that only the control input u_1 depends on time explicitly via the function $h(Z_2, t)$. We will refer to this function as the heat-function, to establish a parallel with well-known probabilistic global minimization methods and underline the primary role of this term in the control, i.e., forcing “motion” as long as the system has not reached the desired equilibrium point, thus preventing the system’s state from converging to other equilibrium points.

According to Proposition 2.1, when one is only interested in the regulation of Z_2 (as in the case of mobile robot path following), any sufficiently regular input $u_1(t)$, which does not asymptotically tend to zero, can be used. This leaves the user with some freedom concerning the choice of this input. For instance, uniform exponential convergence of $\|Z_2(t)\|$ to zero is obtained when $|u_1(t)|$ remains larger than some positive number. Other sufficient conditions for exponential convergence of $\|Z_2(t)\|$ to zero, which do not require $u_1(t)$ to have always the same sign, may also be derived. For example, if $|\dot{u}_1(t)|$ is bounded, it is sufficient to have $|u_1(t)|$ periodically larger than some positive number.

If the application further requires the regulation of z_1 (stabilization of a mobile robot about a fixed desired configuration, for example), then Proposition 2.2 suggests implementing a time-varying feedback u_1 . In both cases, the same control law u_2 (or w_2) based on u_1 -time-scaling can be used.

The conditions imposed by Theorem 2.2 upon the heat-function are not severe and can easily be met. For example, the following three functions

$$h(Z_2, t) = \|Z_2\|^2 \sin(t) \quad (25)$$

$$h(Z_2, t) = \sum_{j=0}^{n-2} a_j \sin(\beta_j t) z_{2+j} \quad (26)$$

$$h(Z_2, t) = \sum_{j=0}^{n-2} a_j \frac{\exp(b_j z_{2+j}) - 1}{\exp(b_j z_{2+j}) + 1} \sin(\beta_j t) \quad (27)$$

(with $a_j \neq 0$, $b_j \neq 0$, $\beta_j \neq 0$, and $\beta_i \neq \beta_j$ when $i \neq j$) satisfy these conditions. For the first function, this is obvious. For the second function, the proof is given in [26]. The same proof basically applies to the third function which presents the additional feature of being uniformly bounded with respect to all its arguments. It can be noted that it is not necessary to use sinusoids at integrally related frequencies, as opposed to the

solution proposed in [34]. In fact, the theorem implies that it is not even necessary to use time-periodic functions, as assumed in most time-varying feedback stabilization results.

For practical purposes, the choice of the heat function is important because the overall control performance (as measured by convergence rate, time needed to enter a small ball centered on zero, sensitivity with respect to perturbations, etc.) critically depends upon this choice. This has been checked by the author in simulation. By performing a complementary analysis in the three-dimensional case, based on Center Manifold techniques, it is also possible to explain why the functions (26) and (27) are better than (25) with respect to the induced asymptotical convergence rate. For the last two functions, the parameters a_j and b_j , which characterize the “slope” of $h(Z_2, t)$ near $Z_2 = 0$ have been found to have much influence on the transient time needed for the system’s solutions to get close to zero. Basically, the larger these parameters are, the shorter the transient time is.

C Stability and Asymptotical Rate of Convergence

From Section II-B, we already know that, when using the smooth time-varying control law (20), (21) with $k_{w_2}(u_1) = k'_{w_2}|u_1|$ ($k'_{w_2} > 0$), the convergence of $\|Z(t)\|$ to zero cannot be exponential. Indeed, $u_1(t)$ would otherwise converge to zero exponentially and the integral $\int_0^t |u_1(\tau)| d\tau$ would not diverge. This would be in contradiction with the fact (pointed out earlier) that divergence of this integral is necessary to the asymptotical convergence of $\|Z_2(t)\|$ to zero.

From the simulation of a smooth time-varying feedback control applied to a unicycle-type vehicle, it has also been observed in [28] that the norm of the state vector did not converge to zero faster than $t^{-\frac{1}{2}}$ for most initial configurations. This is much slower than the uniform exponential rate of convergence that can be obtained in the case of nonlinear systems the linear tangent approximation of which is controllable. It is claimed in [11] that it is not possible to achieve exponential stability for nonholonomic systems by using smooth (differentiable everywhere) time-periodic feedbacks. In mathematical terms, this means that the system’s trajectories cannot satisfy the following inequality

$$\|X(t)\| \leq K \|X(0)\| \exp(-\lambda t) \quad \forall X(0) \text{ in some open ball centered on zero} \quad (28)$$

for some positive real numbers K and λ .

The practical significance of this relation, when it is satisfied, is two-fold: i) the ratio $\|X(t)\|/\|X(0)\|$ between transient and initial errors is uniformly bounded, and ii) all solutions end up converging to zero exponentially.

It is worth noting that these two properties, regrouped under the strong concept of exponential stability, do not necessarily hold together.

For example, the piecewise-continuous time-invariant feedback law proposed by Canudas and Sørvalen in [6], for posture stabilization of a unicycle, only yields the following result

$$\|X(t)\| \leq (K_1 + K_2 \|X(0)\|) \exp(-\lambda t). \quad (29)$$

Each solution converges to zero exponentially, but the slightest initial error, or perturbation, may produce transient deviations the size of which is larger than some constant. Note that this sensitivity to small perturbations, observed for this particular example, is not necessarily indicative of nonsmooth feedbacks as a whole.

In [32], Sørđalen proposes another interesting control which may be seen as a time-varying mix of open-loop and feedback strategies, continuous with respect to time, but nonsmooth in the state vector. He shows that this control, which applies to any chain form system, yields the following property

$$\|X(t)\| \leq g(\|X(0)\|)\exp(-\lambda t) \quad (30)$$

where $g(\cdot)$ is a class \mathcal{K} -function (i.e., strictly increasing and such that $g(0) = 0$) which is not Lipschitz around zero. Precisely, the derivative of $g(x)$ tends to infinity when x tends to zero.

This property has been called \mathcal{K} -exponential stability. It is weaker than the usual exponential stability notion in the sense that the ratio between transient and initial deviations is not uniformly bounded. Nevertheless, it is better than (29) in the sense that the deviations are not lowerbounded by some positive constant. As in the previous case, all solutions tend to zero exponentially in the absence of perturbations acting on the system.

By using the properties of homogeneous systems [12], time-periodic feedbacks which are continuous with respect to both time and state, but not differentiable at the origin, have been proposed in [14], [20], [30] and have been shown to achieve the same type of exponential stability.

Concerning the case of smooth time-varying feedbacks, such as the ones derived in the Section II-B, it is simple to verify that we have

$$\|X(t)\| \leq K\|X(0)\| \quad (31)$$

for some positive constant K the size of which may be taken as close to one as desired via a suitable choice of the control parameters.

Smooth time-varying feedbacks thus are, in some sense, less sensitive to initial errors than the aforementioned nonsmooth feedbacks. The "price" paid for this type of robustness is that the system's solutions do not converge to zero as fast as exponentially. In fact, in view of the above discussion, one can only expect to have

$$\|X(t)\| \leq K\|X(0)\|f(t); \quad f(0) = 1, \quad \lim_{t \rightarrow +\infty} f(t) = 0 \quad (32)$$

where $f(t)$ is a decreasing function which does not tend to zero as fast as exponentially. For example: $f(t) = (1+t)^{-\frac{1}{2}}$, in the case of the control considered in [28], as it may be rigorously established either by applying Center Manifold techniques [7] or by invoking two-time scale techniques, as done in [14].

The purpose of the above discussion was to summarize our actual knowledge concerning the stability properties of

controlled nonholonomic mobile robots (with restricted mobility) in the case of point-stabilization and to point out the difficulty in objectively comparing smooth and nonsmooth feedback solutions. So far, exponential stability, in the usual sense, has not been obtained and is most likely out of reach. Exponential convergence of the solutions to zero, in the absence of perturbations and modeling errors, is possible, however, by using either piecewise-continuous time-invariant feedbacks or continuous time-varying feedbacks which are not differentiable at zero. Smooth time-varying feedbacks are less efficient in terms of asymptotical rate of convergence, but are also potentially not very sensitive to initial conditions and perturbations in the vicinity of zero. A slow asymptotical convergence rate still does not mean that the system's solutions cannot be rapidly steered to an arbitrarily small neighborhood of zero, as pointed out in [34] (for example) and illustrated by simulation results in [29]. Nevertheless, this type of performance has not been obtained for small values of K . This again reflects the apparently unavoidable compromise between performance and robustness. It should also be noted that perturbations acting on nonholonomic systems are not of equal importance depending on the state component which is primarily affected: a deviation in a direction compatible with the vehicle's mobility is clearly not as severe as a deviation which violates one of the system's kinematic constraints (lateral skidding of a car, for example).

Further clarification of these issues is thus needed in relation to a rather fundamental question, seldom addressed in the control literature: Is the asymptotical rate of convergence a good measurement of the overall control performance? Answering this question is not simple, knowing that regulation errors are physically unavoidable and that what often really matters in practice is to keep these errors as small as possible under realistic adverse experimental conditions. While connections between robustness issues and asymptotical rate of convergence of the controlled system have been much studied in the case of linear systems (or nonlinear systems that can be approximated by controllable linear systems), they are still not well understood in other cases.

To illustrate the difficulty with a concrete example, a control law similar to (20)–(21) has been simulated for a three-dimensional unicycle-type vehicle with the following nonsmooth heat-function

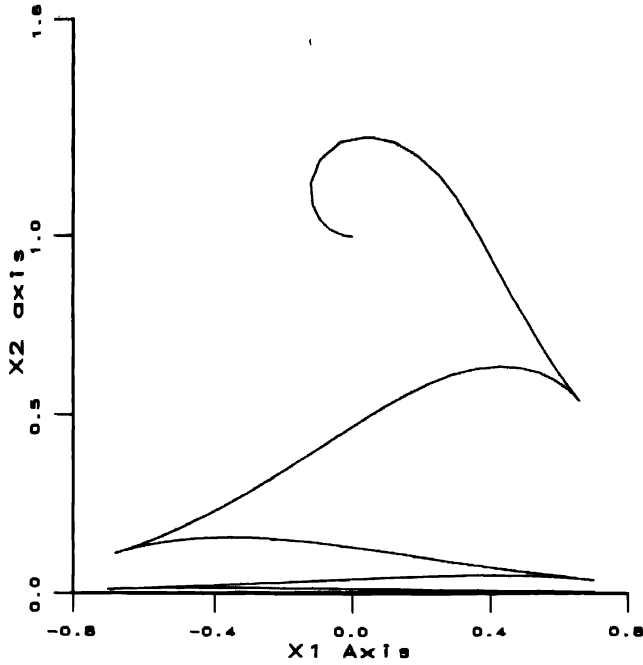
$$h(x_2, x_3, t) = \begin{cases} \sin(t) & \text{if } x_2^2 + (1/k_1)x_3^2 \geq \epsilon^2 \\ 0 & \text{if } x_2^2 + (1/k_1)x_3^2 < \epsilon^2. \end{cases} \quad (33)$$

Note that this function does not satisfy the conditions imposed in Proposition 2.2. The corresponding feedback control is time-varying, but not even continuous.

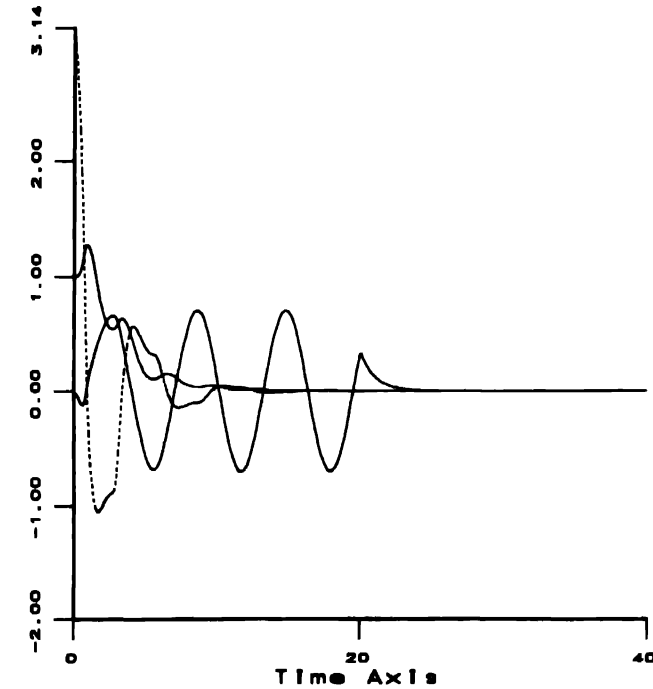
The $(x_1(t), x_2(t))$ Cartesian position of the vehicle is represented in Fig. 1(a).

The time-evolution of $x_1(t)$, $x_2(t)$, and the vehicle's orientation angle $\theta(t)$ ($\approx x_3(t)$) is shown in Fig. 1(b).

After 25 seconds, all variables "seem" to have converged to zero. In reality, it is possible to show that $x_1(t)$ and both control inputs converge to zero exponentially, while



(a)



(b)

Fig. 1.

$2V(X_2(t)) = x_2(t)^2 + (1/k_1)x_3(t)^2$ can only be shown to become smaller than ϵ^2 ($= 10^{-6}$, in the simulation) after a finite time.

This control therefore does not asymptotically stabilize the system to the desired equilibrium, so that the notion of asymptotical rate of convergence does not even apply here. Is this fact sufficient in itself to assert that this is not a good control?

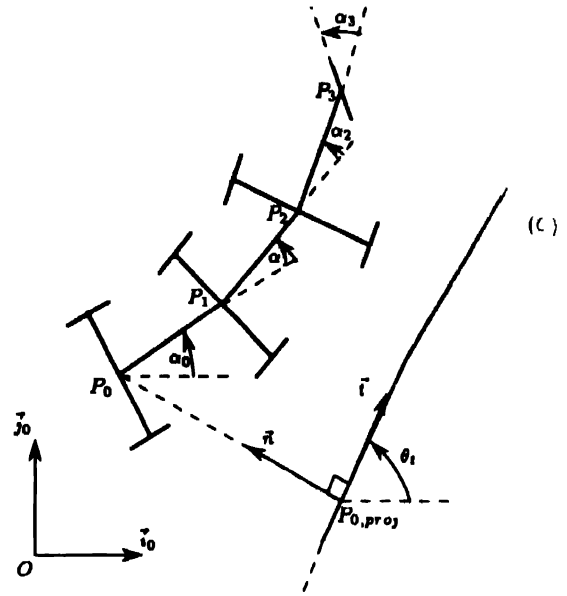


Fig. 2

III. APPLICATION TO THE CONTROL OF A CAR WITH n -TRAILERS

A Modeling Equations and Notations

We consider a car with n -trailers as represented in Fig. 2. The system is assumed to move on flat ground. The wheels are allowed to roll and spin, but not slip.

The vehicle counted first is the last trailer and the following notations are used:

- l_i is the distance between P_i and P_{i+1}
- α_i ($1 \leq i \leq n$) is the angle between $P_{i-1}P_i$ and P_iP_{i+1} which characterizes the orientation of the vehicle $(i+1)$ with respect to the previous vehicle
- α_0 gives the orientation of the first vehicle with respect to a fixed frame. For instance, we may choose: $\alpha_0 = \text{angle}(\vec{v}_0, \vec{P}_0P_1)$
- α_{n+1} is the angle of the car's driving front wheel with respect to the car's body.
- v_i ($0 \leq i \leq n+1$) is the intensity of the velocity of the point P_i . This is the translational velocity of the $(i+1)$ vehicle.
- r is the radius of the car's front steering wheel, and ω the angular velocity of this wheel about its horizontal axis so that $v_{n+1} = r\omega$.

In what follows, only velocity control is considered, and ω and $(d/dt)\alpha_{n+1}$ are chosen as control variables.

Kinematic equations of this system, with respect to a fixed frame, have been derived by various authors. See [13], [22], [35], for example. We will work here with a more general set of equations expressing the system's motion relatively to an arbitrary given path.

A first set of equations is simply obtained by using the classical identity

$$\frac{d}{dt}\vec{P}_{i+1} = \frac{d}{dt}\vec{P}_i + \vec{\omega}_i \wedge \vec{P}_iP_{i+1} \quad \text{for } 0 \leq i \leq n \quad (34)$$

where $\bar{\omega}_i$ is the angular velocity about the vertical axis of the i th vehicle's body.

This yields the following equations

$$\begin{aligned} v_{n+1} &= r\omega \\ v_i &= v_{i+1} \cos(\alpha_{i+1}) \quad (0 \leq i \leq n) \\ \alpha_0 &= v_0^{-1} \frac{\tan(\alpha_1)}{l_1} \\ \dot{\alpha}_i &= v_i \left(\frac{\tan(\alpha_{i+1})}{l_{i+1}} - \frac{\sin(\alpha_i)}{l_i} \right) \quad (1 \leq i \leq n). \end{aligned} \quad (35)$$

We note that the set of variables α_i entirely characterizes the relative positioning of each vehicle with respect to the others.

The remaining equations must describe the motion of one of the vehicles with respect to the path we would like this vehicle to follow. To this purpose, we choose the first vehicle (i.e., the last trailer) and the position coordinates of the point P_0 .

The path to be followed by this point is denoted as (C) . For the sake of simplicity, we consider a smooth simple curve defined by one of its point, the unitary tangent vector at this point, and its curvature $\text{curv}(s)$, with s being the curvilinear coordinate along the curve. Moreover, it is assumed that

- $\text{curv}(s)$ is differentiable $(n+1)$ times. This is necessary for P_0 to be able to remain on (C) without stopping, as it will later appear.
- The radius of any circle tangencing (C) at two or more points and the interior of which does not contain any point of the curve is lowerbounded by some positive real number denoted as r_{\min} . The set of the circles' centers so defined is the Voronoi diagram associated with the curve [33]. This assumption implies in particular that $|\text{curv}(s)| \leq 1/r_{\min}$, $\forall s$. For example, if (C) is a straight line, then $r_{\min} = +\infty$ and $\text{curv}(s) = 0$. If (C) is a circle, then r_{\min} is the circle's radius and $\text{curv}(s) = 1/r_{\min}$.

Under these assumptions, if the distance between P_0 and (C) is smaller than r_{\min} , there is a unique point on (C) , denoted as $P_{0,\text{proj}}$, so that $\|P_0 P_{0,\text{proj}}\|$ is equal to this distance (see Fig. 2).

Let s denote the curvilinear coordinate at the point $P_{0,\text{proj}}$, and $(P_{0,\text{proj}}; \vec{t}, \vec{n})$ the Frenet frame on the curve at this point. The position of P_0 in the plane is completely characterized by the pair of variables (s, y) where y is the intensity of the vector $\vec{P}_{0,\text{proj}} P_0$, i.e.,

$$P_{0,\text{proj}} P_0 = y \vec{n}. \quad (36)$$

Note that in the particular case where (C) is a straight line, s and y coincide with classical Cartesian coordinates. For other curves, one of the control objectives will be to keep the coordinate y smaller than r_{\min} all the time so as to avoid any ambiguity when using the parameterization (s, y) .

This parameterization has previously been proposed in [29] for the control of a unicycle-type vehicle. While it is primarily used here for feedback control purposes, it is also related to the approach developed in [21] for path planning, as shown in the next section.

Let:

- θ_i denote the angle between \vec{v}_0 and $\vec{t}(s)$

- $\theta = \alpha_0 - \theta_i$ the angle between the first vehicle's body and the curve's tangent vector \vec{t} . When the first vehicle follows (C) exactly, with a nonzero translational velocity, θ can only take values equal to $k\pi$ ($k \in \mathbb{Z}$). Without loss of generality one may assume that the desired value for the angle θ is zero.

The following equations for the first vehicle (see also [29]) are then easily derived

$$\begin{aligned} \dot{s} &= v_0 \frac{\cos(\theta)}{1 - \text{curv}(s)y} \\ \dot{y} &= v_0 \sin(\theta) \\ \dot{\theta}_i &= \text{curv}(s) \dot{s} \\ &= v_0 \frac{\text{curv}(s) \cos(\theta)}{1 - \text{curv}(s)y}. \end{aligned} \quad (37)$$

By regrouping (35) and (37) one obtains the following control system

$$\dot{X} = g_1(X) v_0 + g_2 v_2 \quad (38)$$

with

$$X = \begin{bmatrix} s \\ y \\ \theta \\ \alpha_1 \\ \vdots \\ \alpha_{n+1} \end{bmatrix} \quad \dim(X) = n+1$$

$$\begin{aligned} v_0 &= r\omega \prod_{i=4}^{i=n+4} \cos(x_i) \\ v_2 &= \dot{x}_{n+4} \end{aligned}$$

$$g_{1,1}(X) = \frac{\cos(x_4)}{1 - \text{curv}(x_1)x_2}$$

$$g_{1,2}(X) = \sin(x_4)$$

$$g_{1,3}(X) = \frac{\tan(x_4)}{l_1} - \frac{\text{curv}(x_1)\cos(x_3)}{1 - \text{curv}(x_1)x_2}$$

$$g_{1,4}(X) = \frac{1}{\cos(x_4)} \left(\frac{\tan(x_5)}{l_2} - \frac{\sin(x_4)}{l_1} \right)$$

\vdots

$$g_{1,j}(X) = \frac{1}{\prod_{l=4}^{l=j} \cos(x_l)} \left(\frac{\tan(x_{j+1})}{l_{j-2}} - \frac{\sin(x_j)}{l_{j-3}} \right)$$

\vdots

$$g_{1,n+4}(X) = 0$$

$$g_2 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

This control system characterizes our mechanical system, as long as X belongs to the set

$$\Omega = R \times]-r_{\min}, +r_{\min}[\times R \times \left(]-\frac{\pi}{2}, +\frac{\pi}{2}[\right)^{n+1}. \quad (39)$$

Although controllability of the mechanical system in $R^3 \times SO_1^{n+1}$ (i.e., the possibility of steering the system between any two configurations in finite time) has been theoretically established (see [13]), for example, the control design and analysis is hereafter limited to the set Ω . In particular, the angles α_i ($1 \leq i \leq n+1$) are bound to stay in the interval $]-\pi/2, +\pi/2[$. For most practical purposes, this is not restrictive. Moreover, undesirable "jack-knife" effects will systematically be avoided in this way.

B. About Path Planning

As already mentioned, the angle θ must be equal to zero (or π) when P_0 moves along the desired path. This also means that the time-derivative of this angle must be equal to zero. Thus, according to the third equation of the above system, and since y is also equal to zero along the path, one must have

$$\tan(\alpha_1) = \begin{cases} +l_1 \text{curv}(s) & \text{if } \theta = 0 \\ -l_1 \text{curv}(s) & \text{if } \theta = \pi \end{cases} \quad (40)$$

By taking the time derivative of this last equation and comparing the result with the fourth system's equation, one obtains after simple calculations

$$\tan(\alpha_2) = \begin{cases} \frac{l_2}{l_1^2 \text{curv}(s)^2 + 1} \left(\text{curv}(s) + \frac{l_1 \text{curv}^{(1)}(s)}{l_1^2 \text{curv}(s)^2 + 1} \right) & \text{if } \theta = 0 \\ \frac{l_2}{l_1^2 \text{curv}(s)^2 + 1} \left(\text{curv}(s) - \frac{l_1 \text{curv}^{(1)}(s)}{l_1^2 \text{curv}(s)^2 + 1} \right) & \text{if } \theta = \pi \end{cases} \quad (41)$$

where $\text{curv}^{(1)}(s)$ is the first derivative (with respect to the curvilinear coordinate) of the path's curvature.

Repeating the above procedure ($n-1$) times, one would obtain that, along the desired path and for each of the two possible values of θ , the angles α_i ($1 \leq i \leq n+1$) are functions of the path's curvature and its successive derivatives up to the order $(i-1)$. Moreover, one may verify that the correspondence from $]-\pi/2, +\pi/2[^{n+1}$ onto R^{n+1} between the set of angles $\{\alpha_i\}_{1 \leq i \leq n+1}$ and the set $\{\text{curv}^{(i)}(s)\}_{0 \leq i \leq n}$ is one-to-one.

A direct consequence of this fact is that the problem of steering the system between any two configurations (satisfying the aforementioned condition imposed on the range of the angles α_i) can be addressed as a purely geometrical problem consisting of finding a planar path of class C^{n+3} which connects two given points (corresponding to the initial and final position of the point P_0), with given tangents at these points (corresponding to initial and final values of α_0), and conditions imposed on the curvature and its n successive derivatives at both extremities of the path (corresponding to initial and final values of angles α_i ($1 \leq i \leq n+1$)). Since such a path obviously exists, one finds again in this way that the system is controllable in Ω .

The possibility of parameterizing the vehicle's motion by the curvature, and its derivatives, of the path drawn by the first vehicle has been here derived from the system's error equation (38). In [21], this possibility, which yields the above formulation of the path planning problem, is presented as a more general consequence of the system's flatness.

It may also be noted that there is an abundance of methods dealing with this type of geometrical problem. Solution methods proposed in this domain, such as widely used Bezier's polynomial curves (splines) for example [2], could thus be of interest for people working on mobile robot path planning problems and yield methods complementary to those already explored.

C. Smooth Feedback Controls for Path Following and Stabilization about a Fixed Desired Configuration

To apply the results of Section II, system (38) should first be converted into the chain form (1), or, equivalently, to the skew-symmetric chain form (13).

We first rewrite the original system as follows

$$\dot{X} = f_1(X)u_1 + g_2\dot{x}_{n+4} \quad (42)$$

with

$$u_1 = v_0 \frac{\cos(x_3)}{1 - \text{curv}(x_1)x_2}$$

$$\begin{aligned} f_{1,1} &= 1 \\ f_{1,2}(X) &= \frac{1 - \text{curv}(x_1)x_2}{\cos(x_3)} g_{1,2}(x_3) \\ f_{1,3}(X) &= \frac{1 - \text{curv}(x_1)x_2}{\cos(x_3)} g_{1,3}(x_1, x_2, x_3, x_4) \\ f_{1,4}(X) &= \frac{1 - \text{curv}(x_1)x_2}{\cos(x_3)} g_{1,4}(x_1, x_5) \end{aligned}$$

$$\begin{aligned} f_{1,n+3}(X) &= \frac{1 - \text{curv}(x_1)x_2}{\cos(x_3)} g_{1,n+3}(x_4, \dots, x_{n+4}) \\ f_{1,n+4} &= 0. \end{aligned}$$

This control system is equivalent to the original one within the reduced set

$$\Omega_{\text{reduced}} = R \times]-r_{\min}, +r_{\min}[\times \left(]-\frac{\pi}{2}, +\frac{\pi}{2}[\right)^{n+2} \subset \Omega.$$

In particular, due to the choice of the input u_1 , the variable x_3 (i.e., the orientation error θ) has to be kept in the interval $]-\frac{\pi}{2}, +\frac{\pi}{2}[$.

This control system is directly converted into a skew-symmetric chain form via the change of coordinates $\phi_2: X \mapsto Z$, with

$$\begin{aligned} z_1 &= x_1 \\ z_2 &= h_y(x_2) \\ z_3 &= f_{1,2}(x_1, x_2, x_3) \frac{\partial h_y}{\partial x_2} \\ z_4 &= k_1 z_2 + L_{f_1} z_3 \end{aligned}$$

$$z_{j+3} = k_j z_{j+1} + L_{f_1} z_{j+2} \quad (j \leq 2 \leq n+1)$$

(43)

where $h_y(x_2)$ is a smooth monotonic function which maps $]-r_{\min}, +r_{\min}[$ onto R , with first derivative strictly larger than a positive real number, and such that $h_y(0) = 0$. For

example, $h_y(x_2) = (2r_{\min}/\pi)\tan(\pi y/2r_{\min})$ is a possible candidate when $r_{\min} < +\infty$. If $r_{\min} = +\infty$, the simplest choice is $h_y(x_2) = x_2$. This function is introduced here to force $|y(t)| = |x_2(t)|$ to remain smaller than r_{\min} in the subsequent control analysis.

One can verify that the Jacobian matrix $(\partial\phi_2/\partial X)$ is a lower triangular matrix with nonzero components d_i on the diagonal given by

$$D_1 = 1$$

$$d_2 = \frac{\partial h_y}{\partial x_2}$$

$$d_3 = \frac{1 - \text{curv}(x_1)x_2}{\cos(x_3)^2} \frac{\partial h_y}{\partial x_2}$$

$$d_4 = \frac{(1 - \text{curv}(x_1)x_2)^2}{l_1 \cos(x_3)^3 \cos(x_4)^2} \frac{\partial h_y}{\partial x_2}$$

$$d_{n+4} = \frac{(1 - \text{curv}(x_1)x_2)^{n+2}}{(\prod_{i=1}^{i=n+1} l_i)(\prod_{j=3}^{j=n+4} \cos(x_j)^{n+6-j})} \frac{\partial h_y}{\partial x_2}$$

All other components are well defined in Ω_{reduced} . This matrix is thus defined and nonsingular on Ω_{reduced} . Moreover, $\phi_2(\Omega_{\text{reduced}}) = R^{n+4}$. Thus, according to the theorem 0.5 in [23, p. 13], ϕ_2 induces a diffeomorphism of class C^{n+1} ($n+1$ being the degree of differentiability of $\text{curv}(x_1)$) between Ω_{reduced} and R^{n+4} .

Then, by using

$$L_{g_2} L_{f_1}^j z_{i+3} = 0 \quad (0 \leq i \leq n, 0 \leq j \leq n-i) \quad (44)$$

and

$$L_a L_{f_1}^{n+1} z_3 = \frac{(1 - \text{curv}(x_1)x_2)^{n+2}}{(\prod_{i=1}^{i=n+1} l_i)(\prod_{j=3}^{j=n+4} \cos(x_j)^{n+6-j})} \frac{\partial h_y}{\partial x_2}(x_2) \quad (45)$$

it is easy to verify that the control system, with Z as state vector, has the form (13), with the auxiliary control input w_2 defined by

$$w_2 = (k_{n+2}z_{n+3} + L_{f_1} z_{n+4})u_1 + (L_{g_2} L_{f_1}^{n+1} z_3)u_2. \quad (46)$$

Once the system has been put into the skew-symmetric chain form, there only remains to apply Proposition 2.1 to determine a control input w_2 which stabilizes the point $Z_2 = 0$ and thus make the mechanical system follow the path (C) . One can also apply Proposition 2.2 to determine smooth time-varying feedbacks which make the system converge to a given configuration on the path.

Note that stability is only local in this case since the state vector X must belong to Ω_{reduced} . This implies that the angles θ and α_i ($1 \leq n+1$) must have initial values in $]-\frac{\pi}{2}, +\frac{\pi}{2}[$, and that the initial distance $|y(0)|$ must be smaller than r_{\min} .

Remark: The method described above allows asymptotically stabilizing the mechanical system to any configuration so that $|\alpha_i| < (\pi/2)$ ($1 \leq n+1$). In other studies ([22],

[32]), for example, only configurations with zero angles α_i (all trailers aligned) have been considered.

D. A Modification to Broaden the Stability Domain

A practical shortcoming of the above control design method is the necessity of starting with an orientation error $|\theta|$ smaller than $\pi/2$.

This limitation can be removed by considering a more global change of coordinates which converts the initial control system to a modified chain form.

This modification was implicitly used in [29] in the particular case of a unicycle-type vehicle, and simulation results can be found in the same reference.

The new transformation $\phi_3: X \mapsto Z$ that is considered is the following

$$z_1 = x_1$$

$$z_2 = h_y(x_2)$$

$$z_3 = x_3$$

$$z_4 = k_1 \frac{\sin(x_3)}{x_3} \frac{\partial h_y}{\partial x_2}(x_2)z_2 + g_{1,3}(x_1, x_2, x_3, x_4)$$

$$z_j = k_2 z_3 + L_{q_1} z_4$$

$$z_{j+4} = k_{j+1} z_{j+2} + L_{q_1} z_{j+3} \quad (2 \leq j \leq n)$$

(47)

where

- $g_1(X)$ is the vector field involved in the system's representation (38)
- k_j ($1 \leq j \leq n+1$) are positive real numbers
- $h_y(x_2)$ is the monotonic function introduced before.

Remark: Instead of $z_3 = x_3$, one may also take $z_3 = h_\theta(x_3)$, with $h_\theta(x_3)$ being a smooth monotonic function, alike $h_y(x_2)$, which maps an open interval containing $]-\pi, +\pi[$ into R . For example, $h_\theta(x_3) = \tan(kx_3)$ with $k < (1/2)$ can be used. In this case the coordinate z_4 becomes $z_4 = k_1(\sin(x_3)/h_\theta(x_3))(\partial h_y/\partial x_2)(x_2)z_2 + g_{1,3}(x_1, x_2, x_3, x_4)$, and all subsequent coordinates are modified accordingly.

One can verify that $\phi_3(\Omega) = R^{n+4}$ and that the Jacobian matrix $(\partial\phi_3/\partial X)$ is a lower-triangular matrix with nonzero components on the diagonal equal to 1, $(\partial h_y/\partial x_2)(x_2)$, 1, $1/(l_1 \cos(x_4)^2)$, $1/(l_1 l_2 \cos(x_4)^3 \cos(x_5)^2), \dots, 1/((\prod_{i=1}^{i=n+1} l_i)(\prod_{j=4}^{j=n+4} \cos(x_j)^{n+6-j}))$.

The change of coordinates ϕ_3 thus induces a diffeomorphism of class C^{n+1} between Ω and R^{n+4} .

Then, by using

$$L_{g_2} L_{q_1}^j z_{i+3} = 0 \quad (0 \leq i \leq n, 0 \leq j \leq n-i) \quad (48)$$

$$L_{g_2} L_{q_1}^n z_4 = \frac{1}{(\prod_{i=1}^{i=n+1} l_i)(\prod_{j=4}^{j=n+4} \cos(x_j)^{n+6-j})} \quad (49)$$

with the following auxiliary control input w_2

$$w_2 = (k_{n+2}z_{n+3} + L_{g_1} z_{n+4})v_0 + (L_{g_2} L_{q_1}^n z_4)v_2 \quad (50)$$

it is simple to verify that the control system, expressed in terms of the new coordinates, has the following skew-symmetric chain form

$$\begin{aligned} \dot{z}_1 &= v_0 \frac{\cos(z_3)}{1 - \text{curv}(z_1)x_2} \\ \dot{z}_2 &= v_0 \frac{\sin(z_3)}{z_3} \frac{\partial h_y}{\partial x_2}(x_2)z_3 \\ \dot{z}_3 &= -k_1 v_0 \frac{\sin(z_3)}{z_3} \frac{\partial h_y}{\partial x_2}(x_2)z_2 + v_0 z_4 \\ \dot{z}_4 &= -k_2 v_0 z_3 + v_0 z_5 \end{aligned}$$

$$\dot{z}_{j+4} = -k_{j+2} v_0 z_{j+3} + v_0 z_{j+5} \quad (1 \leq j \leq n-1)$$

$$\dot{z}_{n+4} = -k_{n+2} v_0 z_{n+3} + w_2 \quad (51)$$

with $x_2 = h_y^{-1}(z_2)$, and $v_0 = r\omega \prod_{i=1}^{n+4} \cos(x_i)$.

Although this system is not exactly the same as the skew-symmetric chained system (13), a result very similar to Proposition 2.1 can be derived for path following.

Proposition 3.1: If $|v_0(t)|$ and $|\dot{v}_0(t)|$ are bounded, and if the control

$$\begin{aligned} w_2 &= -k_{w_2}(v_0)z_{n+4} \\ (k_{w_2}(\cdot)) &: \text{continuous application strictly positive on } R \setminus \{0\} \end{aligned} \quad (52)$$

is applied to system (51), then the positive function

$$\begin{aligned} V(Z_2) &= 1/2(z_2^2 + (1/k_1)z_3^2 + (1/k_1 k_2)z_4^2 \\ &\quad + \cdots + \left(1/\prod_{j=n+2}^{j=n+4} k_j\right)z_{n+4}^2) \end{aligned} \quad (53)$$

is nonincreasing along the system's solutions, and thus asymptotically converges to some limit value V_{\lim} (which *a priori* depends on the initial conditions).

Moreover, $v_0(t)V(Z_2(t))$ asymptotically converges to zero.

Therefore, if $v_0(t)$ does not converge to zero, then $V_{\lim} = 0$. The submanifold $Z_2 = 0$ is thus globally asymptotically stabilized in this case.

Proof of Proposition 3.1: The proof is quite similar to the proof of Proposition 2.1 except that one has to show at some point that the convergence of $v_0(\sin(z_3)/z_3)(\partial h_y/\partial x_2)(x_2)z_2$ and $v_0 z_3$ to zero yields the convergence of $v_0 z_2$ to zero.

Since $|z_2(t)|$ is bounded (from the boundedness of the Lyapunov function), $(\partial h_y/\partial x_2)(x_2(t))$ is also upperbounded, and $v_0 z_3(\partial h_y/\partial x_2)(x_2)z_2$ thus tends to zero. Therefore, $v_0^2(((\sin(z_3)/z_3))^2 + z_3^2)((\partial h_y/\partial x_2)(x_2))^2 z_2^2)$ also tends to zero.

By assumption, $(\partial h_y/\partial x_2)(x_2)$ is bounded from below by a positive real number. Moreover, the function $((\sin(z_3)/z_3))^2 + z_3^2$ is itself larger than some positive real number. Along a system's solution, it is also bounded from above, since $|z_3(t)|$ is bounded. Using these bounds in the previous convergence result, it is found that $v_0^2 z_2^2$ tends to zero. \diamond

The remarks made after Proposition 2.1 also hold in this case. In particular, adequate values for the control "gains"

k_j ($1 \leq j \leq n+2$) can be determined by comparing, in the neighborhood of $Z_2 = 0$, the control u_2 provided by the proposition and relation (50), with the linearizing feedback (7). Since these gains do not *a priori* depend on the path's shape one may also use, for this comparison, a simpler linear control calculated from the system's tangent linear approximation about the equilibrium ($Z_2 = 0, u_2 = 0$), assuming that the path to be followed is a straight line and that the velocity v_0 is constant.

The problem of stabilizing the system to a fixed desired configuration requires asymptotical convergence of the full state vector Z to zero. A smooth time-varying feedback solution is given in the next complementary proposition.

Proposition 3.2. Consider the same control as in Proposition 3.1

$$w_2 = -k_{w_2}(v_0)z_{n+4} \quad (54)$$

complemented with the following time-varying control

$$v_0 = -k_{v_0} h_s(z_1) + h(Z_2, t) \quad (55)$$

where:

- k_{v_0} is a positive real number.
- $h_s(\cdot)$ is a function of class C^2 which maps R into a bounded interval of R , and such that: i) $h_s(0) = 0$, ii) $0 < h_s^{(1)}(x) < +\infty, \forall x$, and iii) $|h_s^{(2)}(x)| < +\infty, \forall x$. Take for example, the sigmoid function: $h_s(x) = (\exp(ax) - 1/\exp(ax) + 1)$ ($a > 0$).
- $h(Z_2, t)$ is a function with the same properties as in Proposition 2.2.

This control globally asymptotically stabilizes the origin $Z = 0$ of the system (51).

Proof of Theorem 3.2: The first part of the proof consists in showing that $v_0(t)$, and its time derivative are bounded along any system's solution.

Since $\|Z_2(t)\|$ is bounded (due to the boundedness of the Lyapunov function considered in Theorem 3.1), it is clear, from the expression of the control v_0 and the properties of the functions $h_s(z_1)$ and $h(Z_2, t)$, that $|v_0(t)|$ is bounded. As a consequence, $\|\dot{Z}(t)\|$ is also bounded.

Taking the time derivative of the v_0 control law expression

$$\dot{v}_0 = \left[-k_{v_0} h_s^{(1)}(z_1), \frac{\partial h}{\partial Z_2} \right] \dot{Z} + \frac{\partial h}{\partial t} \quad (56)$$

which, in view of the boundedness of $\|Z_2(t)\|$ and $\|\dot{Z}(t)\|$, implies that $|\dot{v}_0(t)|$ is bounded.

Although v_0 is not, strictly speaking, a function of time only (since it is a feedback control), it can be viewed as such along any system's solution, and the results of Proposition 3.1 do apply.

In particular, if $v_0(t)$ does not tend to zero, then $\|Z_2(t)\|$ tends to zero. In this case, $h(Z_2(t), t)$ tends to zero (from condition C_1 and by uniform continuity). From the first system's equation, we also have

$$\dot{z}_1(t) = -k_{v_0} h_s(z_1(t)) + o(t) \quad \text{with } \lim_{t \rightarrow \infty} o(t) = 0. \quad (57)$$

Using the properties of the function $h_s(\cdot)$, the equation implies that the following proposition is true

$$\forall \epsilon > 0, \exists \eta > 0, \exists t_0 : (t > t_0 \text{ and } |z_1(t)| \geq \epsilon) \\ \Rightarrow (z_1(t)\dot{z}_1(t) < -\eta). \quad (58)$$

Since $z_1^2(t)$ cannot remain larger than ϵ^2 with a negative derivative smaller than $-\eta$, there is a time $t_1 (\geq t_0)$ such that $|z_1(t_1)| < \epsilon$. Moreover, after the time t_1 , $|z_1(t)|$ remains smaller than ϵ (since ϵ^2 cannot be reached from below by $z_1^2(t)$ with a negative derivative). The above proposition thus implies

$$\forall \epsilon > 0, \exists t_1 : (t > t_1) \Rightarrow (|z_1(t)| < \epsilon). \quad (59)$$

This is a characterization of the convergence of $z_1(t)$ to zero. In view of the expression of v_0 , this in turn implies that $v_0(t)$ tends to zero (contradiction).

Therefore $v_0(t)$ must tend to zero, implying in turn that $w_2(t)$ and $\dot{z}(t)$ tend to zero. Now, in view of (56)

$$\dot{v}_0(t) = \frac{\partial h}{\partial t}(Z_2(t), t) + o(t).$$

Since $(\partial h / \partial t)(Z_2(t), t)$ is uniformly continuous (its time derivative is bounded), $\dot{v}_0(t)$ tends to zero (Barbalat's lemma). Therefore, $(\partial h / \partial t)(Z_2(t), t)$ also tends to zero. From there, the proof goes on like the proof of Proposition 2.2. \diamond

IV. CONCLUSION

New results about feedback control of chained systems and their application to path following and point stabilization of nonholonomic mobile robots have been presented.

Throughout the paper, feedback control design and analysis have been performed via explicit Lyapunov techniques which apply naturally once the original chain form has been transformed into an equivalent form termed here skew-symmetric chain form. Asymptotical stabilization of the origin of a n -dimensional chained system has been achieved via a two-step approach which allows path following and point stabilization of mobile robots to be treated within the same framework. This approach yields a simple way of determining globally stabilizing smooth time-varying feedbacks the underlying structure of which is easily interpreted. It also applies to the determination of Hölder continuous time-periodic feedbacks which ensure K -exponential stabilization of the origin, as illustrated in [30] in the three-dimensional case. Several smooth and nonsmooth solutions to the point stabilization problem have then been tentatively compared by recalling and commenting upon the type of stability associated with each of them.

Application to path following and point stabilization of a car pulling trailers has been addressed in the second part of the study, based on a specific parameterization of the system's configuration which facilitates the decoupling of the path following problem from translational velocity control and yields general model equations in the chain form. Finally, it has been shown how, by an adequate choice of the system's state coordinates and a slight modification of the original chain form, it is possible to derive feedback control laws endowed with a larger stability domain.

ACKNOWLEDGMENT

The author is grateful to his colleague J.-B. Pomet for his insightful remarks and suggestions.

REFERENCES

- [1] B. d'Andrea-Novet, G. Bastin, and G. Campion, "Modeling and control of nonholonomic wheeled mobile robots," in *Proc. 1991 IEEE Int. Conf. Robotics and Automation*, Sacramento, California, April 1991, pp. 1130-1135.
- [2] P. Bezier, "Courbes et surfaces," *Mathematiques et CAO*, 4, Ed. Hermes, 1986.
- [3] A. M. Bloch and N. H. McClamroch, "Control of mechanical systems with classical nonholonomic constraints," in *Proc. 28th IEEE Conf. Decis. Contr.*, Tampa, FL, 1989, pp. 201-205.
- [4] A. M. Bloch, M. Reyhanoglu, and N. H. McClamroch, "Control and stabilization of nonholonomic dynamic systems," *IEEE Trans. Automat. Contr.*, vol. 37, no. 11, pp. 1746-1757, 1992.
- [5] R. W. Brockett, "Asymptotic stability and feedback stabilization," in *Proc. conf. held at Michigan Technological University, June-July 1982, Progress in Math.*, vol. 27, Birkhauser, pp. 181-208, 1983.
- [6] C. Canudas and O. J. Sørdalen, "Exponential stabilization of mobile robots with nonholonomic constraints," in *Proc. 30th IEEE Conf. Decis. Contr.*, Brighton, UK, Dec. 1991.
- [7] J. Carr, *Applications of Centre Manifold Theory*. New York: Springer-Verlag, 1981.
- [8] J. M. Coron, "Global asymptotic stabilization for controllable systems without drift," *Mathematics of Control, Signals and Systems*. New York: Springer-Verlag, 1992, no. 5, pp. 295-312.
- [9] J. M. Coron, "Links between local controllability and local continuous stabilization," in *Proc. NOLCOS Conf.*, Bordeaux, June 1992, pp. 477-482.
- [10] E. D. Dickmanns and A. Zapp, "Autonomous high speed road vehicle guidance by computer vision," *Preprints 10th IFAC Congress*, Munich, July 1987.
- [11] L. Gurvits and Z. X. Li, "Smooth time-periodic feedback solutions for nonholonomic motion planning," in *Progress in Nonholonomic Motion Planning*, Z. Y. Li and J. Canny, Eds. New York: Kluwer Academic, 1992.
- [12] M. Kowski, "Homogeneous stabilizing feedback laws," *Contr. Theory and Advanced Tech.*, vol. 6, no. 4, pp. 497-516, 1990.
- [13] J. P. Laumond, "Controllability of a multibody mobile robot," in *Proc. Int. Conf. Advanced Robotics, ICAR'91*, Pisa, Italy, June 1991, pp. 1033-1038.
- [14] R. T. M'Closkey and R. Murray, "Convergence rates for nonholonomic systems in power form," in *Proc. 1993 Amer. Contr. Conf.*, San Francisco, 1993.
- [15] A. Miccaelli and G. Samson, "Trajectory tracking for two-steering-wheels mobile robots," in *Proc. Symp. Robot Control'94*, Capri, Sept. 1994.
- [16] S. Monaco and D. Normand-Cyrot, "An introduction to motion planning under multirate digital control," in *Proc. 31st Conf. Decis. Contr.*, Tucson, AZ, 1991, pp. 1780-1785.
- [17] R. M. Murray and S. S. Sastry, "Nonholonomic motion planning: Steering using sinusoids," *IEEE Trans. Automat. Contr.*, vol. 38, no. 5, pp. 700-716, May 1993.
- [18] R. M. Murray, "Applications and extensions of goursat normal form to control of nonlinear systems," in *Proc. 32nd IEEE Conf. Decis. Contr.*, San Antonio, TX, Dec. 1993.
- [19] J. B. Pomet, "Explicit design of time-varying stabilizing control laws for a class of controllable systems without drift," *Syst. Contr. Lett.*, vol. 18, pp. 147-158, 1992.
- [20] J. B. Pomet and C. Samson, "Time-varying exponential stabilization of nonholonomic systems in power form," *INRIA Tech. Rep. 2126*, Dec. 1993.
- [21] P. Rouchon, M. Fliess, J. Levine, and Ph. Martin, "Flatness and motion planning: The car with n -trailers," in *Proc. Int. Conf. ECC'93*, Groningen, Holland, 1993, pp. 1518-1522.
- [22] M. Sampei, T. Tamura, T. Itoh, and M. Nakamichi, "Path tracking control of trailer-like mobile robot," in *Proc. IEEE/RSJ Int. Workshop on Intelligent Robots and Systems '91, IROS'91*, Osaka, Japan, Nov. 1991, pp. 193-198.
- [23] C. Samson, M. Leborgne, and B. Espiau, *Robot Control: The Task-Function Approach* (Oxford Engineering Science Series No. 22). Oxford: Oxford Univ. Press, 1991.
- [24] C. Samson and K. Ait-Abderrahim, "Feedback control of a nonholonomic wheeled cart in Cartesian space," in *Proc. 1991 IEEE Int. Conf. on Robotics and Automation*, Sacramento, CA, Apr. 1991, pp. 1136-1141.

- C. Samson, "Velocity and torque feedback control of a nonholonomic car," in *Proc. Int. Workshop in Adaptive and Nonlinear Control Issues in Robotics*, Grenoble, 1990, France.
- , "Time-varying feedback stabilization of nonholonomic car-like mobile robots," INRIA Tech. Rep. 1515, Sept. 1991.
- , "Time-varying feedback stabilization of car-like wheeled mobile robots," *Int. J. Robotics Research*, vol. 12, no. 1, pp. 55–64, Feb. 1993.
- C. Samson and K. Ait-Abderrahim, "Feedback stabilization of a nonholonomic wheeled mobile robot," in *Proc. IEEE/RSJ Int. Workshop on Intelligent Robots and Systems IROS'91*, Osaka, Japan, Nov. 1991, pp. 1242–1247.
- [29] C. Samson, "Path following and time-varying feedback stabilization of a wheeled mobile robot," in *Proc. Int. Conf. ICARCV 92*, RO-13.1, Singapore, Sept. 1992.
- [30] ———, "Mobile robot control, Part 2: Control of chained systems and application to path following and time-varying point-stabilization of wheeled vehicles," INRIA Tech. Rep. 1594, July 1993.
- [31] O. J. Sørdalen, "Conversion of the kinematics of a car with n trailers into a chained form," in *Proc. 1993 IEEE Int. Conf. Robotics and Automation*, Atlanta, GA, 1993.
- [32] ———, "Feedback control of nonholonomic mobile robots," Ph.D. dissertation, Norwegian Inst. Tech., Trondheim, Norway, March 1993.
- [33] F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*. New York: Springer-Verlag, 1985.
- [34] A. R. Teel, R. M. Murray, and G. Walsh, "Nonholonomic control systems: From steering to stabilization with sinusoids," in *Proc. 31st Conf. Decis. Contr.*, Tucson, AZ, Dec. 1992, pp. 1603–1609.
- [35] D. Tilbury, J.-P. Laumond, R. Murray, S. Sastry, and G. Walsh, "Car-like systems with trailers using sinusoids," in *Proc. IEEE Robotics and Automation*, Nice, France, 1992, pp. 1093–1097.



Claude Samson was born in Buenos Aires, Argentina in 1955. He graduated from the École Supérieure d'Electricité, France, in 1977. He received his Docteur Ingénieur and Docteur d'Etat degrees from the University of Rennes in 1980 and 1983, respectively.

Before joining INRIA (Institut National de Recherche en Automatique et Informatique) in 1981, he spent a year in the Information Systems Laboratory at Stanford University, California, as a visiting scholar. He is presently Directeur de Recherche in the INRIA Centre of Sophia Antipolis and the head of the ICARE Research Group whose activities are centered on robotics and related control issues. His current research interests are in control theory and its applications to the control of mechanical systems, including robot arms, mobile robots, legged mechanisms, underwater vehicles, and satellites.

Dr. Samson is the coauthor with Michel Leboigne and Bernard Espiau of the book *Robot Control: The Task Function Approach* (Oxford University Press, 1991).

Technical Notes and Correspondence

Input-Output Robust Tracking Control Design for Flexible Joint Robots

Zhihua Qu

Abstract—This paper presents the first input-output robust control design for the trajectory following problem of a flexible-joint robot manipulator. The proposed design provides a class of controllers which only require position and velocity feedback and ensure global stability. The resulting stability is that the tracking error can be made to be smaller than a design parameter arbitrarily chosen by the designer. The proposed control is robust since it guarantees tracking performance in the presence of high-order nonlinear uncertainties including unknown joint elasticity, unknown parameters, load variation, and disturbances. Practically, it is more important that no measurement on acceleration, jerk, or position and velocity deformation is required.

I. INTRODUCTION

Control design for robot manipulators with joint flexibility has attracted much attention from control researchers in recent years. The main reason is that, as shown by experimental study in [24], joint flexibility must be taken into account in both modeling and control design to achieve high tracking performance. Common sources of joint flexibility are gear elasticity (for example, harmonic drives), shaft windup, etc. One of the models of flexible joint robots was presented in [20] in which some energy terms are neglected. Nevertheless, most researchers have adopted this approximate model, and we choose to use this model as well in this paper.

Several methods, such as feedback linearization, observer design, adaptive control, singular perturbation, and robust control, have been investigated to design effective control for flexible joint robots. A list of references on these results can be found in the survey paper [22]; we shall only give a brief synopsis of the most recent developments and provide a comparison of different approaches. In feedback linearization method [20, 10], control is designed based on feedback linearization transformation which requires the knowledge of the dynamics and acceleration and jerk measurements. Under the assumption that joint flexibility is "small," there are two time scales in system dynamics, and a control can be designed using singular perturbation technique. Adaptive version of this slow/fast control was studied in [5]–[7]. The resulting control is intuitively simple since it usually contains two parts: a control for rigid body and a corrective control; the control, however, does not guarantee global stability. The most recent adaptive control scheme proposed in [11] guarantees global stability under the standard assumption that dynamics can be fully parameterized. On the front of robust control design, local stability was shown [19] which can be viewed as an extension of robust control results for rigid robots [14], [16]. Later, a robust control guaranteeing global stability was presented in [3]; this full state feedback robust control design is

an application of the result in [18] and an extension of the robust control law [2] for rigid robots. Hybrid control, or combined robust and adaptive control, of flexible joint robots has also been studied [15], [4].

So far all existing control laws for elastic robots require full state feedback. There are some results [12], [25] on designing state observer for flexible joint robots and on the subject of how to close the control loop, that is, how to design a control based on the estimates of the states [13]. The result in [13] combines high-gain observer with high-gain controller to guarantee local stability. Since the separation principle does not hold, in general, for nonlinear systems, it appears unlikely that full state feedback control can be changed to an estimate-based control to guarantee global stability.

In the present paper we consider for the first time the input-output (I/O) control problem of flexible joint manipulators. The control objective is to design a robust control which guarantees global stability and good tracking performance but requires only feedback of link position and velocity. The proposed design method is inspired by the recent result [17] on robust input-output design of mainly linear systems. The proposed control design is not only more attractive in practice since it requires less feedback information, but also allows the presence of significant but bounded nonlinear uncertainties including fast time varying parameters, modeling errors, load change, and unknown flexibility.

The paper is organized as follows. In Section II, dynamics of flexible joint robots and its properties are presented, and necessary assumptions are introduced. In Section III, input-output robust control design is proceeded. Simulation results are presented in Section IV. Finally, some conclusions are made in Section V.

II. PROBLEM FORMULATION

We shall consider the dynamics of a robot with flexible joints to be described by the following nonlinear differential equations

$$0 = M(q_1)\ddot{q}_1 + N(q_1, \dot{q}_1) + K(q_1 - q_2), \quad (1)$$

$$J\ddot{q}_2 = K(q_1 - q_2) - D\dot{q}_2 + \tau + P(q_1, \dot{q}_1) \quad (2)$$

where q_1 and q_2 represent $n \times 1$ vectors of joint angles and motor angles, respectively, τ is $n \times 1$ the control vector of motor output torques, $M(q_1)$ is an $n \times n$ inertia matrix (symmetric and positive definite) for the rigid links, and K is a diagonal matrix representing the joint stiffness. In (1), $N(q_1, \dot{q}_1) = V_m(q_1, \dot{q}_1)\dot{q}_1 + G(q_1) + F_d\dot{q}_1 + F_s(q_1) + T_d$, $V_m(q_1, \dot{q}_1)$ is an $n \times n$ matrix of centripetal and Coriolis terms, $G(q_1)$ is an $n \times 1$ vector of gravity terms, F_d is an $n \times n$ diagonal matrix of dynamic friction coefficients, $F_s(q_1)$ stands for static friction, and T_d denotes the lumped sum of all bounded disturbances and is in general a unknown nonlinear functional of q_1, \dot{q}_1 . In (2), J is a diagonal matrix of actuator inertias reflected to the link side of the gears, D is also a diagonal matrix of torsional damping coefficients, and $P(q_1, \dot{q}_1)$ denotes any possible nonlinearity whose bounding function depends only on q_1 and \dot{q}_1 .

Let q_1^d characterize the desired, smooth trajectory that the robot should track. The assumption of q_1^d being smooth implies that \dot{q}_1^d and

Manuscript received September 10, 1992; revised July 23, 1993 and February 18, 1994. This work is supported in part by U.S. National Science Foundation Grant MSS-9110034.

The author is with the Department of Electrical Engineering, University of Central Florida, Orlando, FL 32826 USA.

IEEE Log Number 9405671.

derivatives up to the second order are continuous and bounded sections of time. The following are the other important properties and assumptions on robot dynamics that will be used in this paper.

- P.1) As shown in [1], the inertia matrix satisfies that, for all $q_1 \in \mathcal{R}^n$, $M I_n \leq M(q_1) \leq \bar{m}(q_1) I_n$ some for known constant \underline{m} and nonnegative function $\bar{m}(q_1)$, where $I_n \in \mathcal{R}^{n \times n}$ is the identity matrix. Function $\bar{m}(q_1)$ reduces to a constant for robots with only revolute joints.
- P.2) It has been shown in [1] that $V_m(q_1, \dot{q}_1)$ is a function of at most of first order in q_1 and \dot{q}_1 . Therefore, there exist constants β_1 and β_2 such that $\|V_m(q_1, \dot{q}_1)\| \leq \beta_1 + \beta_2 \|x_1\| \triangleq \rho_1(x_1)$, where $\|\cdot\|$ denotes Euclidean norm, x_1 is the output vector defined by $x_1 = [e^T \dot{e}^T]^T$, and e is an $n \times 1$ vector representing the trajectory error, i.e., $e = q_1^d - q_1$.
- A.1) Unknown functionals in dynamics equations (1) and (2) include dynamic and static friction, bounded disturbances, etc. It is assumed that they be bounded by known functions as $\|G(q_1) + F_d \dot{q}_1 + F_v(\dot{q}_1)\| \leq \beta_3 + \beta_4 \|x_1\| \triangleq \rho_2(x_1)$, $\|T_d\| \leq \rho_3(x_1, t)$, $\|P(q_1, \dot{q}_1)\| \leq \rho_4(x_1)$, and $\|\dot{M}(q_1)\| \leq \rho_5(x_1)$.
- A.2) The feedback information for control design are only the measurements of joint position and velocity, i.e., q_1 and \dot{q}_1 . The position and velocity deformations $q_2 - q_1$ and $\dot{q}_2 - \dot{q}_1$ are unknown, but initial deformations are bounded, that is, $\|x_2(t_0) - x_1(t_0)\| \leq \rho_6$ or $\|x_2(t_0)\| \leq \rho_7$ for some constant ρ_6 or ρ_7 , where x_2 is the internal state defined by $x_2 = [q_2^T \dot{q}_2^T]^T$.
- A.3) It is physically guaranteed by the property of matrices J , D , $K > 0$ that the fictitious system

$$J\ddot{v} + D\dot{v} + Kv = \ddot{f} + 2\dot{f} + f \quad (3)$$

is uniformly asymptotically stable [23], [8] under zero fictitious input $f(t) = 0$. That is, there exist constants γ_0 , γ_1 , and γ_2 such that the state transition matrix $\Phi(t, t_0)$ associated with (3) satisfies the inequality $\|\Phi(t, t_0)\| \leq \gamma_0 \delta(t) + \gamma_1 e^{-\gamma_2(t-t_0)}$ for all $t \geq t_0$, where $\delta(t)$ is the standard impulse function.

The parameter matrices K , J , and D are unknown and time varying, their values are limited within compact sets as $0 < \underline{K} \leq K \leq \bar{K}$, $0 < \underline{J} \leq J \leq \bar{J}$, $0 < \underline{D} \leq D \leq \bar{D}$, and their derivatives up to the third order are assumed to be bounded.

It is worth making several remarks here about the properties and assumptions listed above. First, for notational simplicity, we consider here the worst situation that no information on dynamics except their bounding functions is known. In many applications in which part of the dynamics are known, only the unknown dynamics need to be bounded since known dynamics can be compensated directly by some nominal (or standard) control law and since only the unknown dynamics have to be bounded and consequently compensated by robust control. The second remark is about Assumption A.3). If the matrices J , D , and K are constants, a lower bound on convergence rate, γ_2 , can be determined from their upper and lower bounds. If the matrices are slowly time varying (that is, time varying parts belong to some space such as L_1 or L_2), γ_2 can also be found by invoking the results in [9]. Only in the case that the matrices are fast time varying, γ_2 has to be assumed since there is no closed-form solution. γ_2 can be estimated in this case, however, because system (3) is always stable. In the event that matrices J , D , and K are known and that there is no uncertainty in (2), the motor angle q_2 can be estimated using linear observer, and I/O control problem reduces to state feedback control problem. The assumption $D > 0$ may always be ensured

physically; although this assumption is needed to show stability under I/O control, the control to be proposed damping and in turn achieves stability of output tracking. This stability imposes no restriction on the value of D . Finally, it will be shown later that Assumption A.2) can be removed so that, in speaking, I/O control becomes globally stabilizing for the system.

The control problem addressed in this paper is to develop an input-output robust control design procedure for flexible joint robots. The main difference between this work and previous work on robust control in [3], [4] is that the control to be designed require only the input and output feedback information, i.e., τ and τ . Measurement of acceleration, or jerk, or deformation is not required. The control is called to be robust since it requires no exact knowledge of the systems, that is, nothing beyond the properties listed above. The control objective is to make the robot track any given trajectory with arbitrarily selected accuracy despite of unknown but bounded nonlinear dynamics.

We can rewrite system dynamics (2) into the following state-space representation

$$\dot{x}_2 = A_2 x_2 + B_2 [\tau + K q_1 + P(q_1, \dot{q}_1)], \quad q_2 = C_2 x_2 \quad (4)$$

where

$$A_2 = \begin{bmatrix} 0 & I_n \\ -J^{-1}K & -J^{-1}D \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 \\ J^{-1} \end{bmatrix},$$

$$C_2 = [I_n \quad 0], \quad x_2 = [q_2^T \dot{q}_2^T]^T.$$

In the general case that the matrices A_2 and B_2 are time varying, the solution to system (4) can be expressed in terms of transition matrix $\Phi(t, t_0)$ defined in Assumption A.3) as

$$q_2(t) = \Phi(t, t_0) * e^{-(t-t_0)} * [\tau + q_1 + P(q_1, \dot{q}_1)] \quad (5)$$

where initial conditions are assumed to be zero, and $*$ denotes the convolution operation.

III. ROBUST I/O CONTROL

To design control without feedback of motor angle q_2 , consider the following linear time-varying system

$$F(s, t)y = v + d \quad (6)$$

where y , v , and d are column vectors of n dimension, y is the output, v is the corresponding input, d is disturbance, and $F(s, t) = Js^2 + Ds + K$ is the so-called time-varying differential operator [23], [28]. That is, system (6) can be rewritten in terms of differential equation as $J\ddot{y} + D\dot{y} + Ky = v + d$.

The intermediate problem to be studied is under what input v the closed-loop system becomes

$$F_o(s)J_o^{-1}J_y y = v + d' \quad (7)$$

where J_o , D_o , and K_o represent any choice of nominal values of J , D , and K respectively, v is any reference input, d' is the equivalent disturbance, and differential operator $F_o(s)$ is given by $F_o(s) = J_o s^2 + D_o s + K_o$. A solution to this problem is provided by the following lemma. Its proof can be found in the Appendix.

Lemma 1: Choose the input v to be

$$v = G_3 y + r + \eta_1 + \eta_2 \quad (8)$$

where G_1 , G_2 , and G_3 are diagonal matrices, and η_1 and η_2 are auxiliary signals defined by

$$\dot{\eta}_1 = -\eta_1 + G_1 v, \quad \dot{\eta}_2 = -\eta_2 + G_2 y. \quad (9)$$

Then, there are unique choices for matrices G_i such that, under the input (8), system (6) is translated to (7) with $d' = (s+1)^{-1}(sI_n + I_n - G_1)d$.

The above lemma provides us a mean of designing I/O robust control. Now, reconsider dynamic equation (2) by rewriting it as

$$\begin{aligned} J\ddot{q}_2 + D\dot{q}_2 + Kq_2 &= \tau + Kq_1 + P(q_1, \dot{q}_1) \\ &= [G_3 q_2 + \tau + \eta_1 + \eta_2] + [Kq_1 + P(q_1, \dot{q}_1) \\ &\quad - G_3 q_2 - \eta_1 - \eta_2] \triangleq v + d \end{aligned}$$

where η_1 and η_2 are given by (9) with $r = \tau$ and $y = q_2$, and G_i are given by the solutions in the proof of Lemma 1 [i.e., (18)]. It follows from the discussion from Lemma 1 that dynamics (2) becomes

$$\ddot{q}_2' + J_o^{-1} D_o \dot{q}_2' + J_o^{-1} K_o q_2' = \tau + P'(q_1, \dot{q}_1, \zeta) \quad (10)$$

where $q_2' = Jq_2$, $\zeta_1 = e^{-t} * \tau$, $\zeta_2 = e^{-t} * \zeta_1$, $\zeta_3 = e^{-t} * \zeta_2$, $\zeta = [\zeta_1^T \zeta_2^T \zeta_3^T]^T$, and

$$\begin{aligned} P'(q_1, \dot{q}_1, \zeta) &= (s+1)^{-1}(sI_n + I_n - G_1) \\ &\quad \cdot [Kq_1 + P(q_1, \dot{q}_1) - G_3 q_2 - \eta_1 - \eta_2]. \end{aligned}$$

For now, the equivalent uncertainty $P'(q_1, \dot{q}_1, \zeta)$ contains the unmeasured state variable q_2 . Later, q_2 will be eliminated using solution (5), and then $P'(q_1, \dot{q}_1, \zeta)$ will be bounded by a known bounding function only of q_1 , \dot{q}_1 , ζ_3 for designing robust control.

It follows from (10) that the solution for q_2 is given by

$$\begin{aligned} q_2(t) &= J^{-1} C_2 e^{A_{2o} t} J_o' x_2'(t_0) + J^{-1} \bar{\tau} \\ &\quad + J^{-1} C_2 e^{A_{2o} t} B_{2o} J_o P'(q_1, \dot{q}_1, \zeta) \quad (11) \end{aligned}$$

where $x_2'(t_0) = [q_2^T(t_0)J(t_0) \dot{q}_2^T J(t_0) + q_2^T(t_0)\dot{J}(t_0)]^T$, A_{2o} and B_{2o} are the matrices A_2 and B_2 , respectively, with $J = J_o$, $K = K_o$, and $D = D_o$, and $\bar{\tau}$, a filtered version of τ , is defined and can be calculated by

$$\bar{\tau} = C_2 z, \quad z = [z_1^T \ z_2^T]^T, \quad \dot{z} = A_{2o} z + B_{2o} J_o \tau. \quad (12)$$

Substituting solution (11) into (1) yields the output error system

$$\dot{x}_1 = A_1 x_1 + B_1 (\Delta A - J^{-1} \bar{\tau} + q_1) \quad (13)$$

where

$$A_1 = \begin{bmatrix} 0 & I_n \\ -M^{-1}(q_1)K_p & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 \\ M^{-1}(q_1)K \end{bmatrix}$$

$\Delta A = K^{-1}K_p e_1 + K^{-1}M(q_1)\ddot{q}_1^T + K^{-1}N(q_1, \dot{q}_1) - J^{-1}C_2 e^{A_{2o} t} x_2'(t_0) - J^{-1}C_2 e^{A_{2o} t} B_{2o} J_o P'(q_1, \dot{q}_1, \zeta)$, and $K_p > 0$ is a diagonal gain matrix chosen by the designer.

Stability analysis and control design will be done using Lyapunov's direct method. To apply the Lyapunov technique, let us choose

Lyapunov function candidate, a positive definite function with respect to x_1 and $z = u$, to be

$$V = V_1 + V_2, \quad V_1 = \frac{1}{2} e_1^T K_p e_1 + \frac{1}{2} (\alpha e_1 + \dot{e}_1)^T M(q_1) (\alpha e_1 + \dot{e}_1),$$

$$V_2 = \frac{1}{2} (z - u)^T (z - u) \quad (14)$$

where u is the vector to be chosen later, and $\alpha > 0$ is a constant scalar. The following lemma illustrates the property of sub-Lyapunov function $V_1(x_1)$ along the trajectory of subsystem (13), and its proof is included in Appendix.

Lemma 2: Consider systems (1) and (2) under Properties P.1) and P.2) and Assumptions A.1), A.2), and A.3). The time derivative of the sub-Lyapunov function $V_1(x_1)$ satisfies the following inequality along the trajectory of system (13)

$$\dot{V}_1(x_1) \leq -\lambda_1 \|e_1\|^2 + \|w(x_1)\| \varphi_1(x_1, \zeta_3) - w^T(x_1) K J^{-1} \bar{\tau} \quad (15)$$

where $\lambda_1 = \alpha \lambda_{\min}(K_p)$, $\lambda_{\min}(\cdot)$, and $\lambda_{\max}(\cdot)$ represent the operation of taking the minimum and maximum eigenvalues, respectively, $w(x_1) \triangleq [\alpha I_n \ I_n] x_1 = \alpha e_1 + \dot{e}_1$, and $\varphi_1(x_1, \zeta_3)$ is a well-defined function only of x_1 and ζ_3 .

Note that bounding function $\varphi_1(x_1, \zeta_3)$ does not depend on internal state x_2 and only requires an estimate (or bound) on the initial condition of x_2 . In $\varphi_1(x_1, \zeta_3)$, only one of bounding terms on initial conditions is included and is denoted by that associated with ρ_7 . In the rest of the proofs of Lemmas 1 and 2, we choose to leave the terms of initial conditions out of bounding functions for simplicity, and we can remove the term containing ρ_7 as well. Although the system is highly nonlinear, this treatment is legitimate since nonzero initial conditions obtained in input-output modeling come from exponentially stable linear systems such as (11) and (9). Consequently, nonzero initial conditions contribute in \dot{V} only exponentially decaying time functions multiplied by terms that are linearly proportional in \dot{V} to the norm of the state. Note that Lyapunov function and its time derivative still have quadratic-like expressions. Thus, the terms associated with nonzero initial conditions can be combined with negative definite quadratic terms using the inequality $-a^2 + 2|a||b| \leq b^2$ to conclude the exactly same stability result. That is, Assumption A.2) can be removed, and nonzero initial conditions need not be considered in stability analysis. The trade-off of assuming zero initial conditions is that the terms bounding initial conditions may converge slow and therefore dominate the behavior of the overall system. By compensating nonzero initial conditions, convergence rate is what is designed and found in stability proof.

Upon having Lemma 1, we can proceed input-output robust control design, the proposed I/O controller is generated by the following recursive mapping

$$\begin{aligned} u &= [u_1^T \ u_2^T]^T, \quad u_1 = u_{11} + u_{12}, \quad u_2 = u_{21} + u_{22}, \\ u_{11} &= \alpha \bar{J} \bar{K}^{-1} \bar{m}(q_1) w(z_2), \\ u_{12} &= \bar{J} \bar{K}^{-1} \frac{\mu(x_1, \zeta_3, t) \|\mu_1(x_1, \zeta_3, t)\|^{v_1}}{\|\mu_1(x_1, \zeta_3, t)\|^{v_1+1} + \epsilon_1^{1+v_1}} \varphi_1(x_1, \zeta_3), \\ u_{21} &= \alpha(u_1 - z_1), \\ u_{22} &= \frac{\mu_2(x_1, z_1, \zeta_2, \zeta_3, t) \|\mu_2(x_1, z_1, \zeta_2, \zeta_3, t)\|^{v_2}}{\|\mu_2(x_1, z_1, \zeta_2, \zeta_3, t)\|^{v_2+1} + \epsilon_2^{1+v_2}} \\ &\quad \cdot g_2(x_1, z_1, \zeta_2, \zeta_3, t), \\ \tau &= \tau_1 + \tau_2, \quad \tau_1 = \alpha u_2 + u_1 + (J_o^{-1} K_o - I_n) z_1 \\ &\quad + (J_o^{-1} D_o - \alpha I_n) z_2, \\ \tau_2 &= \frac{\mu_3(x_1, z, \zeta, t)}{\|\mu_3(x_1, z, \zeta, t)\| + \epsilon_3} g_3(x_1, z, \zeta, t) \quad (16) \end{aligned}$$

here $\epsilon_k > 0$ and $\nu_j \geq 0$, $k = 1, 2, 3$ and $j = 1, 2$, are constants, the constants ν_j are chosen such that the first-order partial derivatives of u_{j2} with respect to their variables are well defined, and $\mu_1(x_1, \zeta_1, t) = w(x_1)\varphi_1(x_1, \zeta_1)$, $g_2(x_1, z_1, \zeta_2, \zeta_3, t) = \|\dot{u}_1\| + \|\bar{J}\|^{-1}\|\bar{K}\|$, $\mu_2(x_1, z_1, \zeta_2, \zeta_3, t) = (u_1 - z_1)g_2(x_1, z_1, \zeta_2, \zeta_3, t)$, $g_3(x_1, z_1, \zeta, t) = \|\dot{u}_2\|$, $\mu_3(x_1, z_1, \zeta, t) = (u_2 - z_2)g_3(x_1, z_1, \zeta, t)$.

The recursive mapping basically involves finding bounding functions of $\|\dot{u}_1\|$ and $\|\dot{u}_2\|$ for obtaining u_2 and τ , respectively. These bounding functions should not be functions of the state derivatives but only of the system state. To this end, the terms $\|\dot{u}_1\|$ and $\|\dot{u}_2\|$ can be bounded by first developing bounds for the first-order partial derivatives of u_i with respect to its variables and then by determining the bounds for the first-order time derivatives of its variables using dynamic equations (12) and (13). The bounding functions are usually obtained by taking the Euclidean norm. This design process yields that, since u_1 is a function of x_1 and ζ_1 , $g_2(\cdot)$ and therefore u_2 are functions of x_1 , ζ_2 , ζ_3 and z_1 , and that $g_3(\cdot)$ and therefore τ are functions of x_1 , ζ and z . As a result, the control design procedure is well defined. There have been several ways to guarantee differentiability of robust control [17]. In the case where there are known dynamics in the system, u_{11} should be modified to compensate. Similarly, u_{21} and τ_1 can be redesigned since not every term in \dot{u}_1 and \dot{u}_2 is uncertain. For simplicity, we leave the details to the reader.

We are now in a position to state the following theorem on control (16) which requires only input-output measurements and is robust with respect to any bounded parameters and uncertainties for manipulators with flexible joints. Its proof can be found in the Appendix.

Theorem: Consider system (1) and (2) under Properties P.1) and P.2) and Assumptions A.1), A.2), and A.3). If I/O robust controller τ is chosen to be the outcome of the mapping procedure (16), then, while all internal state variables are uniformly bounded, the trajectory tracking error x_1 (both e and \dot{e}) is globally and uniformly ultimately bounded. That is, as time approaches infinity, the magnitude of the tracking error characterized by $V_1(e, \dot{e})$ becomes no larger than the design parameter ϵ^* where

$$\epsilon^* = \frac{1}{2\alpha \min \left\{ \frac{\lambda_{\min}(K_p)}{\lambda_{\max}(K_p)}, 1 \right\}} \sum_{i=1}^l \epsilon_i.$$

Furthermore, the robust control τ is continuous and globally, uniformly bounded.

It should be noted that, for good transient performance under nonzero initial conditions, ϵ_i should be chosen to be proportional to the initial condition of measurable sub-state, $x_1(t_0)$. In this case, the stability of uniform ultimate boundedness becomes Lyapunov stability, and the stability is global since $x_1(t_0)$ is available as part of state feedback information. Under zero conditions, ϵ_i should be chosen to be any small constant.

IV. SIMULATION

To illustrate the effectiveness of the proposed input-output robust control scheme, consider a two-link flexible joint robot. An I/O robust controller will be designed for the robot using the nonlinear mapping (16). Each step in mapping (16) requires to find a bounding function for the time derivative of the previous fictitious controller and, if necessary, this derivation procedure can be automated using symbolic manipulation package. Note that a sophisticated nonlinear function of some variables can always be bounded by a simpler but

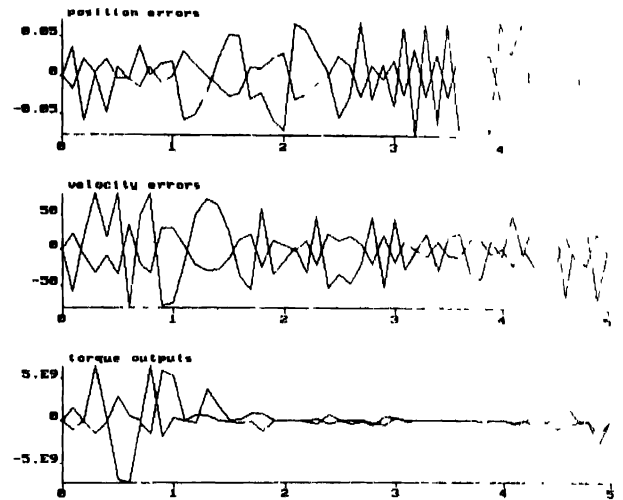


Fig. 1. Simulation with stiffness $K(t) = 100 + 90 \sin(0.5t)$

higher order polynomial functions of the same variables. Since the proposed control has complete robustness, even a control with oversimplified expressions works very well. Based on this observation, we choose robust controllers u_1 , u_2 , and τ in the form of (16) as follows: $\varphi_1 = (1 + \|x_1\|^2)/10$, $\bar{m}(q_1) = 2$, $\epsilon_1 = \epsilon_2 = \epsilon_3 = 1.0$, $\alpha = 35$, $K_p = I$, $\bar{J}K^{-1} = 5I$, $\nu_1 = \nu_2 = 2$, $D_o = 0.5I$, $K_o = 90I$, $J_o = I$, $g_2 = [1 + (\varphi_1)^2 + \|u_1 - z_1\|^2 + \|w\|^2]/10$, $g_3 = [1 + g_2^2 + \|u_2 - z_2\|^2]/10$. The auxiliary state z is defined by (12) using the nominal parameters J_o , D_o , and K_o .

For a two-link flexible joint robot, the expressions of $M(q)$ and $N(q, \dot{q})$ in the dynamic equation (1) can be found in [1]. For the simulation purpose, we made the following choices: $m_1 = m_2 = 1$, $I_1 = I_2 = 1$, $F_{d1} = F_{d2} = 0$, $J = I$, $D = I$, $K(t) = 100 + 90 \sin(0.5t)$, $T_d(j) = \sin(0.2q_1(j)) + 0.1 \cos(t)$, and $P(q_1, \dot{q}_1)(j) = \cos(\|\dot{q}_1\|^2 + \dot{q}_2^2(j))$, where $j = 1, 2$. The desired trajectory for joint angle to track is $q^d = [1 - \cos(t)]I$.

The robust control was simulated using SIMNON[®], and the simulation results are shown in the figures. Further research is needed to determine bounding functions such that robust control requires less control magnitude.

V. CONCLUSION

A robust control design method was presented in this paper for robot manipulators with flexible joints. In contrast to existing robust control results and to other design techniques such as adaptive control and singular perturbation, the proposed control guarantees global stability while using only measurement of joint position and velocity. Because acceleration and jerk measurement is no longer required, the proposed control is more practical and attractive. Furthermore, the proposed control works robustly to compensate any unknown but bounded uncertainties including unknown flexibility, (time-varying) parameters, load variation, and disturbances. Regardless of uncertainties, position and velocity tracking errors have been shown to satisfy any arbitrarily small tracking accuracy requirement. In return of achieving such a complete robustness, the designer is required off-line to perform differentiation and to determine bounding functions in order to find robust controller through the proposed nonlinear mapping. The off-line calculation can be automated using a software package with the capability of symbolic manipulations.

APPENDIX

Proof of Lemma 1: Utilizing differential operator, we can rewrite the input v as

$$\begin{aligned} v &= G_3 y + r + (s+1)^{-1} [G_1 v + G_2 y] \\ &= (sI_n + I_n - G_1)^{-1} \{[(s+1)G_3 + G_2]y + (s+1)r\}. \end{aligned}$$

Substituting the above equation into (6) yields

$$\begin{aligned} [(sI_n + I_n - G_1)(Js^2 + Ds + K) - (s+1)G_3 - G_2]y \\ = (s+1)r + (sI_n + I_n - G_1)d. \end{aligned}$$

It then follows that the closed-loop system is given by (7) with $d' = (s+1)^{-1}(sI_n + I_n - G_1)d$ iff the following identity holds

$$\begin{aligned} [(sI_n + I_n - G_1)(Js^2 + Ds + K) - (s+1)G_3 - G_2] \\ = (s+1)(J_o s^2 + D_o s + K_o)J_o^{-1}J. \quad (17) \end{aligned}$$

Since all matrices are diagonal, the above equality can be rewritten as

$$\begin{aligned} [(s+1-G_{1i})(J_i s^2 + D_i s + K_i) - (s+1)G_{3i} - G_{2i}] \\ = (s+1)(J_{oi} s^2 + D_{oi} s + K_{oi}) \frac{J_i}{J_o} \end{aligned}$$

where subscript i denotes the i th element on the diagonal. It is easy to verify that the above equality holds if G_1 , G_2 , and G_3 are chosen such that

$$\begin{aligned} G_{1i} &= \frac{D_i}{J_i} - 2\frac{\dot{J}_i}{J_i} - \frac{D_{oi}}{J_{oi}}, \\ G_{3i} &= \dot{D}_i + K_i + (1-G_{1i})D_i - 3\dot{J}_i - 2\ddot{J}_i \\ &\quad - 2\frac{D_{oi}}{J_{oi}}\dot{J}_i - \frac{K_{oi}}{J_{oi}}J_i - \frac{D_{oi}}{J_{oi}}J_i, \\ G_{2i} &= \dot{K}_i + (1-G_{1i})\dot{K}_i - \dot{G}_{1i} - \frac{d^3 J_i}{dt^3} \\ &\quad - \left(1 + \frac{D_{oi}}{J_{oi}}\right)\ddot{J}_i - \frac{K_{oi} + D_{oi}}{J_{oi}}\dot{J}_i - \frac{K_{oi}}{J_{oi}}J_i. \quad (18) \end{aligned}$$

Solution (18) guarantees identity (17) which in turn guarantees that the closed-loop system is internally stable. The above relations will be used to determine known bounds on G_1 , G_2 , and G_3 . These bounds are required in designing robust control. \square

Proof of Lemma 2. It follows from (13) and (14) that

$$\begin{aligned} \dot{V}_1(x_1) &= e_1^T K_p \dot{e}_1 + \frac{1}{2} w^T(x_1) \dot{M}(q_1) w(x_1) \\ &\quad + w^T(x_1) \dot{M}(q_1) (\alpha \dot{e}_1 + \dot{e}_1) \\ &= -\alpha e_1^T K_p e_1 - w^T(x_1) K J^{-1} \bar{\tau} + u^T(x_1) \{0.5 \dot{M}(q_1) u(x_1) \\ &\quad + \alpha \dot{M}(q_1) \dot{e}_1 + K \Delta A + K q_1\}. \end{aligned}$$

Thus, the proof can be completed if the uncertainty $0.5 \dot{M}(q_1) w(x_1) + \alpha \dot{M}(q_1) \dot{e}_1 + K \Delta A + K q_1$ can be bounded in euclidean norm by a well-defined function $\varphi_1(q_1, \dot{q}_1, \zeta_1)$.

It follows from (9) and (8) that $(sI_n + I_n - G_1)\eta_1 = G_1 G_3 q_2 + G_1 \tau + G_1 \eta_2$ and $(s+1)\eta_2 = G_2 q_2$. Hence, we have

$$\begin{aligned} P'(q_1, \dot{q}_1, \zeta_1) &= K q_1 + P(q_1, \dot{q}_1) - G_3 q_2 - (s+1)^{-1} \\ &\quad \cdot [G_1 K q_1 + G_1 P(q_1, \dot{q}_1) + G_2 q_2] - (s+1)^{-1} G_1 \tau \\ &= K q_1 + P(q_1, \dot{q}_1) - (s+1) G_3 \frac{1}{s+1} q_2 + \dot{G}_3 \frac{1}{s+1} q_2 \\ &\quad - (s+1)^{-1} [G_1 K q_1 + G_1 P(q_1, \dot{q}_1)] - G_2 \frac{1}{s+1} q_2 \\ &\quad + \frac{1}{s+1} \dot{G}_2 \frac{1}{s+1} q_2 - G_1 \zeta_1 + (s+1)^{-1} \dot{G}_1 \zeta_1 \\ &= K q_1 + P(q_1, \dot{q}_1) - (s+1) G_3 \frac{1}{s+1} q_2 + \dot{G}_3 \frac{1}{s+1} q_2 \end{aligned}$$

$$\begin{aligned} &- (s+1)^{-1} [G_1 K q_1 + G_1 P(q_1, \dot{q}_1)] - G_2 \frac{1}{s+1} q_2 \\ &+ \frac{1}{s+1} \dot{G}_2 \frac{1}{s+1} q_2 \\ &- (s+1)^{-2} [G_1 \zeta_1 - 3(s+1)^{-1} \dot{G}_1 \zeta_1 + 3(s+1)^{-2} \\ &\quad \cdot \ddot{G}_1 \zeta_1 - (s+1)^{-3} \frac{d^3 G_1}{dt^3} \zeta_1] \end{aligned}$$

in which the last two steps are derived by integration by part. Such steps are necessary to find a well-defined robust control τ by guaranteeing that the second-order time derivative of the bounding function to be determined does not depend on τ explicitly.

The bounding function $\varphi_1(x_1, \zeta_1)$ that bounds uncertainties in \dot{V}_1 can then be chosen to be

$$\begin{aligned} \varphi_1(x_1, \zeta_1) &= \|\bar{K}\| \|q_1\| + \|K_p e_1\| + \overline{m}(q_1) \|\ddot{q}_1\| + \rho_1(x_1) \|\dot{q}_1\| \\ &\quad + \rho_2(x_1) + \rho_3(x_1) + \frac{1}{2} \|w(x_1) \rho_5(x_1) + \alpha \overline{m}(q_1) \|\dot{e}_1\| \\ &\quad + \|\bar{K}\| \|\underline{J}\|^{-1} \|\zeta_2 e^{-\lambda_2(t-t_0)}\| \cdot \|(\bar{J} + \|\bar{J}\|)\| \rho_7 + \|G_1\| \|\cdot\| \|\zeta_1\| \\ &\quad + \int_{t_0}^t \|C_2 e^{-\lambda_2 o(t-s)} B_{2o} J_o\| \|\|\bar{K}\| \|q_1\| + \rho_4(x_1) \\ &\quad + (\|\dot{G}_3\| + \|G_2\|) \\ &\quad \cdot \left| \frac{1}{s+1} q_2 \right| + 3 \|\ddot{G}_1\| \|\cdot\| \|\zeta_1\| \Big] ds \\ &\quad + \int_{t_0}^t \|C_2' e^{-\lambda_2 o(t-s)} B_{2o} J_o\| \|\|\zeta_1\| \\ &\quad \cdot \|\zeta_1\| ds + \int_{t_0}^t \|C_2'' e^{-\lambda_2 o(t-s)} B_{2o} J_o\| (3 \|\dot{G}_1\| \|\cdot\| \|\zeta_1\| \\ &\quad + \|\dot{G}_3\|) \cdot \left| \frac{1}{s+1} q_2 \right| ds \\ &\quad + \int_{t_0}^t \|C_3 e^{-\lambda_3(t-s)} B_1\| (\|\bar{K}\| \|\cdot\| \|G_1\| \|q_1\| + \|\dot{G}_1\| \\ &\quad \cdot \rho_4(x_1) + \|\dot{G}_2\| \|\cdot\| \left| \frac{1}{s+1} q_2 \right| + \left| \frac{d^3 G_1}{dt^3} \right| \|\cdot\| \|\zeta_1\|) ds \Big\} \end{aligned}$$

where $\|\cdot\|$ denotes a known bound or bounding function on the norm of the argument

$$A_1 = \begin{bmatrix} 0 & I_n & 0 \\ 0 & 0 & I_n \\ -J_o^{-1} K_o & -J_o^{-1} (K_o + D_o) & -I_n - J_o^{-1} D_o \end{bmatrix},$$

$$C_2' = [I_n - J_o^{-1} K_o \quad 2I_n - J_o^{-1} D_o],$$

$$C_2'' = [I_n \quad I_n], \quad B_1 = [0 \quad 0 \quad I_n]^T.$$

$$C_3 = [I_n \quad 0 \quad 0]$$

and \bar{J} denotes the upper bound on \dot{J} . Although q_2 is not measured, a bounding function of measured variables can be determined from (5) for $\|(1/s+1)q_2\|$ as

$$\begin{aligned} \left\| \frac{1}{s+1} q_2 \right\| &\leq (\gamma_0 \delta(t) + \gamma_1 e^{-\gamma_2(t-t_0)}) \\ &\quad \cdot \left[\|\zeta_1\| + e^{-(t-t_0)} * e^{-(t-t_0)} * (\|q_1\| + \rho_4(x_1)) \right] \triangleq \left| \frac{1}{s+1} q_2 \right|. \end{aligned}$$

Hence, $\varphi_1(x_1, \zeta_1)$ is a well-defined function only of x_1 and ζ_1 . \square

Proof of Theorem Taking time derivative of V_2 along the trajectory of the system given by (12) yields $\dot{V}_2 = (\bar{\tau} - u_1)^T (\dot{\tau}_1 - \dot{\tau}_2 - u_2)^T (-J^{-1}K_{\tau-1} - J^{-1}D_{\tau-2} + \tau - u_2)$. It follows from $V_1 + V_2$ and from Lemma 2 that

$$\begin{aligned} & \lambda_1 \|e_1\|^2 + \|u(t_1)\| \varphi_1(\tau_1 - \zeta_1) - u^T(t_1) K J^{-1} \bar{\tau} \\ & + (\bar{\tau} - u_1)^T (\dot{\tau}_2 - u_2) \\ & + (\dot{\tau}_2 - u_2)^T (-J^{-1}K_{\tau-1} - J^{-1}D_{\tau-2} + \tau - u_2) \\ & - \lambda_1 \|e_1\|^2 + \|u(t_1)\| \varphi_1(\tau_1 - \zeta_1) - u^T(t_1) K J^{-1} u_1 \\ & + (\tau_1 - u_1)^T [u_2 - u_1 - J^{-1}K u(t_1)] \\ & + (\dot{\tau}_2 - u_2)^T (\tau_1 - u_1 - J^{-1}K_{\tau-1} - J^{-1}D_{\tau-2} + \tau - u_2) \\ & \leq -\lambda_1 \|e_1\|^2 - \alpha \bar{m}(q_1) \|u(t_1)\|^2 - \alpha (\tau_1 - u_1)^2 \\ & - \alpha (\dot{\tau}_2 - u_2)^2 \\ & + \|u(t_1)\| \varphi_1(\tau_1) - u^T(t_1) K J^{-1} u_1 + (\tau_1 - u_1)^T u_2 \\ & + \|\tau_1 - u_1\| \\ & \quad \left[\|u(t_1)\| + \left| K^T u(t_1) \right| \right] + (\dot{\tau}_2 - u_2)^T \tau_1 + \|\dot{\tau}_2 - u_2\| \|u(t_1)\| \\ & \leq -\lambda_1 \|e_1\|^2 - \alpha \bar{m}(q_1) \|u(t_1)\|^2 \\ & - \alpha (\tau_1 - u_1)^2 - \alpha (\dot{\tau}_2 - u_2)^2 + \epsilon_1 + \epsilon_2 + \epsilon_3 \\ & \leq -\lambda^* V + \lambda^* \epsilon^* \end{aligned}$$

where $\lambda^* = \min \{2\lambda_1/\lambda, \lambda_1(K_{\tau-1} - 2\alpha)\}$ and $\epsilon^* = (\epsilon_1 + \epsilon_2 + \epsilon_3)/\lambda^*$. Solving the above first order differential equation yields

$$V(t) \leq e^{-\lambda^*(t-t_1)} V(t_1) + \epsilon^*(1 - e^{-\lambda^*(t-t_1)}) \rightarrow \epsilon^* \quad \text{as } t \rightarrow \infty.$$

Therefore V and the output tracking error e_1 (i.e. τ and ϵ) are uniformly ultimately bounded and their convergence is exponential.

The uniform boundedness of all internal variables can be established as follows. First V_2 being bounded implies that both $\tau_1 - u_1$ and $\dot{\tau}_2 - u_2$ are uniformly bounded. Second it follows from e_1 being bounded that q_1 is bounded and hence u_1 is uniformly bounded. It follows from (16) that $\|u_1\| \leq \bar{K}^{-1} \varphi_1(\tau_1 - \zeta_1)$ which shows that u_1 is uniformly bounded if ζ_1 is bounded. It follows from (92) that

$$\dot{\zeta}_1 = \frac{1}{(\gamma+1)} [Iq_1 + Dq_1 - P(q_1 - q_1)]$$

Note that τ_1 (and therefore q_1 and \dot{q}_1) is bounded. Even if matrices I, D, P are time varying we can show using integration by part (as did in the proof of Lemma 3) that ζ_1 is a function of signal $(1/\gamma+1)q_2$ and its filtered versions through stable transfer functions. On the other hand it follows from (1) that

$$\frac{1}{\gamma+1} \dot{q}_2 = \frac{1}{\gamma+1} \{K^{-1} [M(q_1)q_1 + \lambda(q_1 - q_1) + Kq_1]\}$$

The terms on the right hand side of the above equation can be shown again using integration by part to be functions of q_1, \dot{q}_1 and their filtered versions. Thus we conclude from the above two equations that ζ_1 and u_1 are globally and uniformly bounded. Now one can conclude boundedness of τ_1 since u_1 is bounded. The boundedness of τ_1 then implies boundedness of ζ_1 which in turn shows boundedness of u and therefore τ . Finally boundedness of τ_2 is used to establish in turn boundedness of $\zeta_1, \tau_1, \tau_2, \tau$ and τ_2 . \square

REFERENCES

[1] J. J. Craig, *Adaptive Control of Mechanical Manipulators*, Reading, MA: Addison-Wesley, 1988.
[2] D. M. Dawson, Z. Qu, F. L. Lewis, and J. F. Dorsey, "Robust control for the tracking of robot motion," *Int. J. Contr.*, vol. 52, no. 3, pp. 581-595, Sep. 1990.

[3] D. M. Dawson, Z. Qu, M. Bridges, and J. C. Carroll, "Control of rigid link flexible joint electrically driven robots," in *Conf. Decis. Contr.*, Brighton, UK, Dec. 1991, pp. 1-5.
[4] D. Dawson, Z. Qu, and M. Bridges, "Hybrid adaptive tracking of rigid link flexible joint robots," in *Proc. 34th Annual Meeting of the IEEE Conf. on Decision and Control*, pp. 95-98, Atlanta, GA, Dec. 1991, pp. 1-5.
[5] F. Ghorbel, J. Y. Hung, and M. W. Spong, "Adaptive flexible joint manipulators," in *Proc. 1989 IEEE Int. Conf. on Automation*, Scottsdale, Arizona, 1989, pp. 15-19.
[6] F. Ghorbel and M. W. Spong, "Stability analysis of adaptively controlled flexible joint manipulators," in *Proc. 29th IEEE Conf. Decis. Contr.*, Honolulu, HI, 1990, pp. 2538-2544.
[7] ———, "Adaptive integral manifold control of flexible joint robot manipulators," in *Proc. IEEE Conf. Robotics and Automation*, Nice, France, 1992, pp. 707-714.
[8] A. Ilchmann, I. Nurnberger, and W. Schmalk, "Time varying polynomial matrix systems," *Int. J. Contr.*, vol. 40, no. 2, pp. 329-362, 1984.
[9] H. Khalil, *Nonlinear Systems*, New York: MacMillan, 1992.
[10] K. Khorasani, "Nonlinear feedback control of flexible joint manipulators: A single link case study," *IEEE Trans. Automat. Contr.*, vol. 35, no. 10, pp. 1145-1149, 1990.
[11] R. Lozano and B. Brogliato, "Adaptive control of robot manipulators with flexible joints," *IEEE Trans. Automat. Contr.*, vol. 37, no. 2, pp. 174-181, 1992.
[12] S. Nicotia, P. Tomei, and A. Tornambe, "A nonlinear observer for elastic robots," *IEEE Trans. Robotics and Automation*, vol. 4, no. 1, pp. 45-52, 1988.
[13] S. Nicotia and P. Tomei, "State observers for rigid and elastic joint robots," *Robotics and Computer Integrated Manufacturing*, vol. 9, no. 2, pp. 113-120, 1992.
[14] Z. Qu, J. F. Dorsey, X. Zhang, and D. M. Dawson, "Robust control of robots by computed torque law," *Syst. Contr. Lett.*, vol. 16, no. 1, pp. 25-32, 1991.
[15] Z. Qu, D. M. Dawson, and J. F. Dorsey, "Exponentially stable trajectory following of robotic manipulators under a class of adaptive controls," *Automatica*, vol. 28, no. 3, pp. 579-586, May 1992.
[16] Z. Qu and J. F. Dorsey, "Robust tracking control of robots by a linear feedback law," *IEEE Trans. Automat. Contr.*, vol. 36, no. 9, pp. 1081-1084, Sep. 1991.
[17] Z. Qu and D. M. Dawson, "Model reference robust control of a class of SISO systems," in *Proc. 1992 Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 1182-1186.
[18] ———, "Lyapunov direct design of robust tracking control for classes of cascaded nonlinear uncertain systems without matching conditions," in *Proc. 30th IEEE Conf. Decis. Contr.*, Brighton, UK, Dec. 1991, pp. 2521-2526.
[19] Z. Qu, "Robust control of a class of nonlinear uncertain systems with applications to flexible joint robots," *IEEE Trans. Automat. Contr.*, vol. 37, no. 9, pp. 1437-1442, Sep. 1992.
[20] M. W. Spong, "Modeling and control of elastic joint robots," *J. Dynamic Syst. Measurement Contr.*, vol. 109, pp. 310-319, 1987.
[21] ———, "Adaptive control of flexible joint manipulators," *Syst. Contr. Lett.*, vol. 13, pp. 15-21, 1989.
[22] ———, "The control of flexible joint robots: A survey," in *New Trends and Applications of Distributed Parameter Control Systems* (Lecture Notes in Pure and Applied Mathematics), G. Chen, F. B. Lee, W. Littman, and L. Markus, Eds., New York: Marcel Dekker, 1990.
[23] A. V. Solodov, *Linear Automatic Control Systems with Varying Parameters*, New York: American Elsevier, 1966.
[24] I. M. Sweet and M. C. Good, "Redefinition of the robot motion control problem: effects of plant dynamics, drive system constraints, and user requirement," in *Proc. 23rd IEEE Conf. Decis. Contr.*, Las Vegas, NV, 1984.
[25] P. Tomei, "An observer for flexible joint robots," *IEEE Trans. Automat. Contr.*, vol. 35, no. 6, pp. 739-743, 1990.

The Use of Symbolic Computation in Nonlinear Control: Is It Viable?

Bram de Jager

Abstract—To help along the analysis and design of nonlinear control systems the NONLINCON package, an acronym for *Nonlinear Control*, was developed. This note addresses the usefulness of symbolic computation, and of the NONLINCON package in particular, for the symbolic analysis and design of nonlinear control systems. The symbolic computation program MAPLE is used as computing substratum. Textbook problems show that the NONLINCON package can be used successfully. A larger scale problem is too complex, however, to be solved with the current versions of NONLINCON and MAPLE. The conclusion is that symbolic computation is a viable approach for textbook problems, but not yet for more complex ones.

I. INTRODUCTION

The use of symbolic computation for control purposes is investigated by several researchers. Zeitz *et al.*, [1] applies the program MACNON, based on MACSYMA, to analyze observability and reachability, and to design observers and controllers for nonlinear systems. Blankenship [2] also used MACSYMA to solve some control problems with his implementation CONDENS. The use of MAPLE for several control problems is reported in [3]. Some problems reported in this note, e.g., with solving partial differential equations, are partly resolved in [4]. They describe a MAPLE package, here called NonCon, that can compute, e.g., the zero dynamics and provide solutions for exact linearization problems. In this note we illustrate the use of this package by applying it to some textbook problems and a more complex one.

The main goals and contributions of this note are

- a proof of the viability of symbolic computation for some problems in the analysis and design of nonlinear control systems
- to show the characteristics of a prototype implementation
- to give some examples and to document some applications
- to discuss directions for future research
- to familiarize a larger audience in the control community with the use of symbolic computation.

The note is structured as follows. First, Section II formulates the control problems and presents the algorithms used. The implementation in NONLINCON of these algorithms is treated in Section III. Section IV presents textbook examples for some problem areas. A more complex example is treated in Section V. Section VI closes with conclusions and gives directions for future research.

II. THE PROBLEMS

Of several areas in nonlinear control, where symbolic computation is likely to be of some profit, we discuss the computation of the normal form, the zero dynamics, and the input-output and state-space exact linearization. In the presentation of these problems, we follow Isidori [5].

A. Preliminaries

We start with a nonlinear model of a plant and assume that it can be described adequately by a set of nonlinear differential equations,

affine in the input u , and without direct feed-through from input to output

$$\dot{x} = f(x) + g(x)u, \quad y = h(x) \quad (1)$$

with state vector $x \in \mathbb{R}^n$, containing all necessary information of the plant, input vector $u \in \mathbb{R}^m$, and output vector $y \in \mathbb{R}^m$. The number of inputs is equal to the number of outputs, i.e., the plant is square. This assumption is for convenience only. Part of the theory can also be derived if the number of inputs is not equal to the number of outputs. The vector field f is smooth, g has m columns g_i of smooth vector fields, and h is a column of m scalar-valued smooth functions h_i .

The type of control law chosen is static state-feedback. Therefore, the value of the input vector $u(t)$ depends on the state $x(t)$ and a new reference input vector $v(t)$ as

$$u = \alpha(x) + \beta(x)v \quad (2)$$

where the components α_i and β_{ij} are smooth functions.

For nonlinear models it is appropriate to allow for a nonlinear change of coordinates

$$z = \Phi(x). \quad (3)$$

It is required that the Jacobian $\partial\Phi/\partial x$ of the transformation Φ is, at least locally, invertible for Φ to qualify as a change of coordinates.

B. Relative Degree

The nonlinear model (1) is said to have a vector relative degree $\{r_1, \dots, r_m\}$ at $x = x''$ if

- 1) $L_{g_j} L_f^{k_j} h_i(x) = 0$ for $k_j = 1, \dots, r_j - 2$ ($i, j = 1, \dots, m$) and all x in a neighborhood of x'' ,
- 2) the following $m \times m$ matrix is nonsingular at x''

$$A(x) = \begin{bmatrix} L_{g_1} L_f^{r_1-1} h_1(x) & \dots & L_{g_m} L_f^{r_m-1} h_1(x) \\ \vdots & & \vdots \\ L_{g_1} L_f^{r_1} h_m(x) & \dots & L_{g_m} L_f^{r_m} h_m(x) \end{bmatrix}$$

Here $L_f^k h_i(x)$ means the k th successive Lie derivative of the scalar function $h_i(x)$ in the direction of the vector field f , e.g., $L_f h_i(x) = (\partial h_i(x)/\partial x)f(x)$. The matrix A is sometimes called the decoupling matrix. The relative degree can also be interpreted as the number of times the outputs have to be differentiated before the input explicitly appears.

The models we consider do not necessarily have a relative degree, either because the first condition cannot be satisfied or because the matrix A is singular at x'' .

C. Normal Form

When the model has a well-defined relative degree we can use a change of coordinates (3), with $z = (\xi, \eta)$, to transform (1) to the normal form

$$\begin{aligned} y_i &= h_i(x) = \xi_i^1 \\ \dot{\xi}_i^1 &= \xi_i^2 \\ &\vdots \\ \dot{\xi}_i^{r_i} &= b_i(\xi, \eta) + \sum_{j=1}^{r_i} a_{ij}(\xi, \eta) u_j \quad \text{for } i = 1, \dots, m \\ \dot{\eta}_i &= q_i(\xi, \eta) + p_i(\xi, \eta) u \quad \text{for } i = r+1, \dots, n \end{aligned} \quad (4)$$

Manuscript received March 26, 1993; revised February 12, 1994.

The author is with the Faculty of Mechanical Engineering, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands. IEEE Log Number 9405674.

with $r = \sum_{j=1}^m r_j$ and

$$a_{ij}(\xi, \eta) = L_{q_j} L_f^{r_j-1} h_i(\Phi^{-1}(\xi, \eta))$$

$$h_i(\xi, \eta) = L_f^{r_i} h_i(\Phi^{-1}(\xi, \eta)) \quad \text{for } i, j = 1, \dots, m.$$

The terms a_{ij} are the entries of A and we can compactly write [with only equations containing the input u in (4)]

$$\dot{\xi}^{(r)} = b(\xi, \eta) + A(\xi, \eta)u \quad (5)$$

$$\dot{\eta} = q(\xi, \eta) + p(\xi, \eta)u$$

where $\xi^{(r)}$ contains $\dot{\xi}_i^{(r)}$, $i = 1, \dots, m$. The equation for $\dot{\eta}$ is called the internal dynamics. Because A is nonsingular if the relative degree is well defined, the control

$$u = A^{-1}(v - b) \quad (6)$$

with the new input v , is properly defined and linearizes the part of model (5) that is visible at the output.

D. Zero Dynamics

The zero dynamics problem is: obtain the dynamics of the model when the output y is required to be zero for all t , by a proper choice of initial state $x(0)$ and control input $u^*(t)$. Here we have to employ an appropriate static state feedback and use proper initial conditions. More specific: we are looking for the locally maximal output zeroing submanifold and its associated dynamics.

When the model has a well-defined relative degree, the zero dynamics follows from the normal form, by substitution of the output nulling input u^* and using the property that the states ξ can be set to zero in the internal dynamics.

For models without a relative degree the zero dynamics can be computed by using the Zero Dynamics Algorithm. The way this algorithm works is by considering a sequence of nested submanifolds M_i , with $M_i \supset M_{i+1}$ and $M_0 = h^{-1}(0)$, i.e., the first submanifold is the inverse image of the point $y = 0$. When some conditions are fulfilled this sequence converges to the locally maximal output zeroing submanifold Z^* in some neighborhood of x'' and there exists a mapping u^* such that $f^*(x) = f(x) + g(x)u^*(x)$ is tangent to Z^* . The pair (Z^*, f^*) is called the zero dynamics of the model. When the mapping $H(x)$ is defined in a neighborhood U of x'' by $Z^* \cap U = \{x \in U; H(x) = 0\}$ the input u^* can be computed as the solution of $L_f H(x) + L_{g_j} H(x)u^* = 0$. For further details of this algorithm and the conditions to be imposed to guarantee convergence of the sequence to Z^* , we refer to [5, Section 6.1].

E. Input-Output Exact Linearization

The input-output exact linearization problem is: find out if it is possible to transform model (1) to a linear one by state feedback (2) and compute this feedback. The linearity property should be established between the new input v and the output y . Formally, we are looking for a neighborhood U of x'' and a static state feedback such that for all $k \geq 0$ and all $1 \leq i, j \leq m$ the expression $L_{(q_j)} L_f^k h_i(x)$ is independent of x on U .

For models with a well-defined relative degree, the input-output exact linearizing feedback follows from the normal form and is given by (6).

For models without a relative degree, the static state feedback that solves the problem can be computed with the Structure Algorithm. This algorithm uses a sequence of Toeplitz matrices

$$M_k(x) = \begin{bmatrix} T_0(x) & \cdots & T_k(x) \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T_0(x) \end{bmatrix}$$

where

$$T_k(x) = \begin{bmatrix} L_{q_1} L_f^k h_1(x) & \cdots & L_{q_m} L_f^k h_m(x) \\ \vdots & \ddots & \vdots \\ L_{q_1} L_f^k h_1(x) & \cdots & L_{q_m} L_f^k h_m(x) \end{bmatrix}$$

for $0 \leq k \leq 2n-1$. The conditions for existence can be cast in rank conditions on $M_k(x)$. Within the Structure Algorithm, feedback functions $\alpha(x)$ and $\beta(x)$ are constructed. For further details of this algorithm we refer to [5, Section 5.4].

F. State-Space Exact Linearization

The state-space exact linearization problem is: under which conditions is it possible to transform the model (1) to a linear and controllable one by state feedback (2) and a change of coordinates (3)? The linearity property should be established between the new input v and the transformed state z .

This problem has been solved. The solution is only valid for models with a well-defined relative degree and requires the existence of (synthetic) outputs for which the model has a full order relative degree, $r = n$. When this cannot be obtained, it is sometimes convenient to strive after a maximal relative degree. Then the corresponding input-output linearizing state feedback realizes a minimal dimension of the internal dynamics.

Using the Lie product

$$[f, g_i] = \frac{\partial g_i}{\partial x} f - \frac{\partial f}{\partial x} g_i$$

define the adjoint $\text{ad}_f^k g_i$ recursively as $\text{ad}_f^k g_i = [f, \text{ad}_f^{k-1} g_i]$ with $\text{ad}_f^0 g_i = g_i$. Then define the distributions

$$G_i = \text{span} \{ \text{ad}_f^k g_j; 0 \leq k \leq i, 1 \leq j \leq m \} \quad \text{for } 0 \leq i \leq n-1.$$

We now state the conditions for a solution [5, Theorem 5.2.4].

Theorem 1. Given the model

$$\dot{x} = f(x) + g(x)u, \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m$$

with $\text{rank } g(x'') = m$. There exists a solution for the state-space exact linearization problem if and only if

- 1) G_i has constant dimension near x'' for each $0 \leq i \leq n-1$
- 2) G_{n-1} has dimension n
- 3) G_i is involutive for each $0 \leq i \leq n-2$.

Here, involutive means that the distribution is closed under the Lie product, i.e., the dimension of the distribution G_i does not change when a vector field, generated by the Lie product of each combination of two of the vector fields in G_i , is added to the distribution.

When the given conditions are fulfilled, there exist solutions $\lambda_i(x)$, $i = 1, \dots, m$, for the following partial differential equations

$$L_{g_j} L_f^k \lambda_i(x) = 0, \quad \text{for } 0 \leq k \leq r_i - 2 \quad \text{and} \quad 1 \leq j \leq m.$$

Also $\sum_{i=1}^m r_i = n$, where the set of integers $\{r_1, \dots, r_m\}$ is the relative degree vector. The m functions λ_i can be computed, based on a constructive proof of Theorem 1. Using the functions λ_i , the change of coordinates Φ and state feedback $u = \alpha(x) + \beta(x)v$ follow.

III. SOLUTION WITH SYMBOLIC COMPUTATION

An analysis of the algorithms shows that symbolic computation programs should be able to compute the Lie derivative and Lie product, the Jacobian, the rank, the Gauss Jordan decomposition, the inverse of a matrix, and the determinant of a matrix; to do matrix-vector and matrix-matrix multiplication; and to test involutiveness. More mundane facilities like symbolic substitution are needed also. The main problems are the computation of solutions for sets of nonlinear equations, e.g., to compute the kernel of a mapping, and

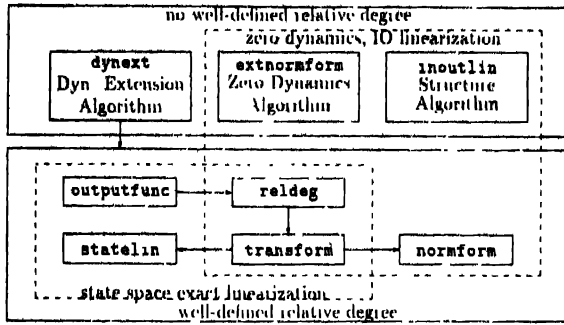


Fig. 1. Structure of NONLINCON.

for sets of partial differential equations. These problems present the major bottlenecks for the symbolic solution of our problems.

Solutions for the problems described in the previous section

- 1) the normal form
- 2) the zero dynamics
- 3) the input-output exact linearization
- 4) the state-space exact linearization

are included in NONLINCON. For problems 1) and 4) the model should have a well-defined relative degree; for 2) and 3) this is not necessary. Implementations of other algorithms, e.g., the Dynamic Extension Algorithm and a method to solve partial differential equations, are included in this package also. The structure of the implementation is sketched in Fig. 1. For a thorough discussion of the implementation, see [6].

Above we noted that a main problem was to compute solutions for sets of nonlinear equations. In MAPLE a facility is available to solve sets of nonlinear equations, namely the `solve` function, but this facility is not always sufficient. We did not try to improve this.

We also noted that another main problem was to solve sets of partial differential equations. In MAPLE almost no facilities are available for their solution. To improve this, the following route was chosen.

The partial differential equations we have to solve are from the "completely integrable" type; in other words, based on Frobenius' Theorem, we know that a solution for the partial differential equations exists. To compute the solutions Frobenius' Theorem itself is of no help. This problem was solved by computing the solutions with an algorithm based on a constructive proof of Frobenius' Theorem. The procedure is as follows. The solution of the partial differential equation can be constructed from the solutions of related sets of ordinary differential equations. Because MAPLE provides some facilities for this type of problems, the `dsolve` command, the problem seems solved. The `dsolve` command is not very powerful, however, and is often unable to present a solution, although this solution is known to exist. Therefore the `dsolve` procedure was extended in an ad hoc way, to handle a larger class of problems, by setting up a recursive procedure to solve sets of differential equations, starting from the "shortest" (assumed to be the simplest) equation, substituting the solution in the remaining equations, and so on. No effort was spent in trying to detect a (block) triangular dependency structure in the set of differential equations, which would be a more rigorous option. Despite this extension the solution of the partial differential equations is often unsuccessful, so NONLINCON cannot complete the computations.

A minor problem was that some standard MAPLE `linalg` functions are only suitable for rational polynomials. This was too limited for our purposes. Therefore, the rank and implicitly the `gausselim` and `gaussjord` procedures were extended, so a larger class of problems could be handled. This resulted in the new functions `extrank`, `extgausselim`, and `extgaussjord`.

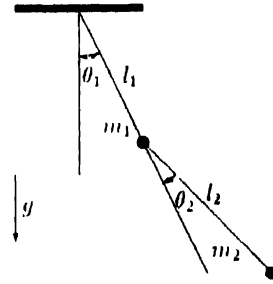


Fig. 2. Robot with two revolute joints.

IV. TEXTBOOK PROBLEMS

To illustrate the use of NONLINCON we consider several examples. The first example computes the zero dynamics of a single-input single-output model with a well-defined relative degree. For the second example the zero dynamics is computed with the Zero Dynamics Algorithm. For the last example the input-output linearizing state feedback is computed with the Structure Algorithm. The last two examples are based on models that do not have a well-defined relative degree. All examples are contrived ones. The first example is taken from [7, Example 12.43] and the other two from [5, Examples 6.1.2 and 5.4.1].

Example 1: The model for a robot with two revolute joints (see Fig. 2) can be derived from the kinetic and potential energy T and V

$$2T = m_1 l_1^2 \dot{\theta}_1^2 + m_2 (l_1^2 \dot{\theta}_1^2 + l_2^2 (\dot{\theta}_1 + \dot{\theta}_2)^2 + 2l_1 l_2 (\dot{\theta}_1 + \dot{\theta}_2) \cos \theta_2)$$

$$V = -g \{ (m_1 + m_2) l_1 \cos \theta_1 + m_2 l_2 \cos (\theta_1 + \theta_2) \}.$$

The inputs u_1 and u_2 are the joint torques and the outputs y_1 and y_2 the joint positions. The state x of the model corresponds with the degrees-of-freedom and their derivatives in the following way $x^T = [\theta_1 \ \theta_2 \ \dot{\theta}_1 \ \dot{\theta}_2]$. The aim is to derive the dynamics of the model when it is constrained. There are two cases, the first one with $y = y_1 = \theta_1$ constrained to 0 and $u_2 = 0$, the second one with $y = y_2 = \theta_2$ constrained to 0 and $u_1 = 0$. According to [7] the zero dynamics are given by (with all model parameters set to one)

$$\dot{x}_2 = x_4 \quad \dot{x}_4 = g \sin x_2$$

respectively

$$\dot{x}_1 = x_3 \quad \dot{x}_3 = (3/5)g \sin x_1.$$

The following log of a NONLINCON session shows that these results can be reproduced. Two functions are used: `normform` to compute the zero dynamics and `transform` for the inverse transformation. The log for the first case, where $y = \theta_1$, is found at the bottom of the next page (as marked by (x)). The second case for $y = \theta_2$ is found at the bottom of the next page, as marked by (y). The results coincide with the ones given above.

Example 2: The model of the system is

$$\dot{x} = \begin{bmatrix} x_2 \\ x_4 \\ x_1 x_4 \\ x_5 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ x_3 & x_2 \\ 0 & 1 \\ x_5 & x_2 \\ 1 & 1 \end{bmatrix} u, \quad y = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

This model has no well-defined relative degree at $x'' = 0$ because

$$A = \begin{bmatrix} 1 & 0 \\ x_3 & x_2 \end{bmatrix}$$

singular for $\dot{x}_2 = 0$. The zero dynamics is given in [5] by $\dot{x}_1 = -x_1$ and the zeroing input by

$$\begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

The next edited log of a NonLinCon session shows that the zero dynamics can be computed. From the functions supplied we only needed `extnoimform` that implements the Zero Dynamics Algorithm as shown at the bottom of the page (as indicated by (7)). The term computed last (for \dot{x}_1) represents the zero dynamics. The results coincide with the ones given above.

Example 3 The model of the system is

$$\begin{aligned} \dot{x}_1 + \dot{x}_2 \\ \dot{x}_1 \dot{x}_2 \\ -x_1 + x_2 + u = 0 \\ + x_1^2 \end{aligned}$$

This model has no well defined relative degree because

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

is singular for all t . According to [5] the necessary feedback for input-output linearization is $u = a(t) + b(t)\dot{x}_1$ with

$$a(t) = -x_1 + x_2, \quad b(t) = 2x_1 + 1$$

The following shows that NonLinCon can compute the input-output linearizing state feedback. From the functions supplied we only need `outputlin` that implements the Structure Algorithm which is found at the bottom of the next page. The computed state feedback agrees with the previous result.

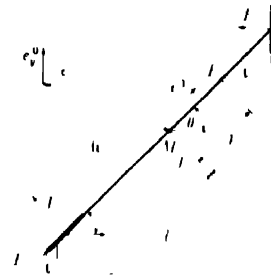


Fig. 3 One track model of a vehicle

V. VEHICLE SIMULATION PROBLEM

This more complex problem is derived from an inverse simulation problem in multibody dynamics. A vehicle is required to perform a standardized maneuver. To simulate the maneuver the inputs to the model must be known. Normally the maneuver is such that a suitable selected output of the model can be set to zero. Often this output does not fully determine the behavior of the model and the remaining freedom represents exactly the zero dynamics. If this dynamics is stable then the simulation is also stable; if it is unstable then additional measures are needed for a stable simulation.

As an example we use a simple two dimensional one track model of a vehicle with traction and cornering forces acting on the tires. It looks like the model of a bike because it only considers the center line of a motor vehicle (see Fig. 3).

The required maneuver is steady state turning: the longitudinal speed of the center of mass M is constant and a point P on the

```
* normform *
% dynamic 1
[eta[1]dot - eta[2], eta[2]dot - grav*sin(eta[1])]
output nulling input 1
2
[ -sin(eta[1]) eta[2] + cos(eta[1]) grav*sin(eta[1]) ]
* transform *
inverse transformation
{x[1] eta[1], x[4] eta[2], x[3] zeta[1], x[5] zeta[2]}

* normform *
% zero dynamic 1
[eta[1]dot - eta[2], eta[2]dot - 3/5*grav*sin(eta[1])]
output nulling input 1 [ 1/5*grav*sin(eta[1]) ]
* transform *
inverse transformation
{x[4] zeta[2], x[3] eta[2], x[2]-zeta[1], x[1] eta[1]}

* extnoimform *
1st step, output zeroing submanifold
[x[2] - 0, x[1] - 0]
2nd step, output zeroing submanifold
[x[2] - 0, x[1] - 0, x[4] - 0]
3rd step, output zeroing submanifold
[x[2] - 0, x[1] - 0, x[4] - 0, x[5] - 0]
3rd step, matrix L_gH has full rank m
the zero dynamics (zerodyn)
{x[1]dot-0, x[2]dot-0, x[3]dot--x[3], x[4]dot-0, x[5]dot-0}
the zeroing input (uzero) [ 0, - [3] ]
```

(x)

(y)

(z)

center line should describe a circle of specified radius. The specific problem is for which distances p , from P to M , the zero-dynamics is stable or unstable.

The equations of motion of the model are

$$\begin{aligned} m\ddot{x}_m &= -F_f \sin(\delta + \theta) - F_r \sin \theta + F_{t_r} \cos \theta \\ m\ddot{y}_m &= +F_f \cos(\delta + \theta) + F_r \cos \theta + F_{t_r} \sin \theta \\ J\ddot{\theta} &= aF_f \cos \delta - bF_r \end{aligned}$$

with inputs F_{t_r} (traction force) and δ (steering angle). The three degrees-of-freedom x_m , y_m , and θ , are, respectively, the coordinates of M and the orientation of the vehicle with respect to a fixed reference frame (\vec{e}_x^0 , \vec{e}_y^0), indicated by the superscript 0. The steering angle δ and the drift angles α_f and α_r are given with respect to a body fixed reference frame. The vehicle has mass m and moment of inertia J with respect to M . The forces acting on the vehicle are the traction force F_{t_r} and the lateral tire forces F_f , F_r . The last two forces can be expressed in the drift angle and the normal tire force F_n by the so-called magic formula for pure slip [8]

$$F(F_n, \alpha) = D(F_n) \sin(C \operatorname{atan}(B\xi - E(B\xi - \operatorname{atan}(B\xi)))) \quad (7)$$

with $\xi = \alpha + S_h$. The dependency holds for both front and rear. The parameters in this formula have to be fitted to experimental data. The drift angles can be expressed in the degrees-of-freedom and steering angle by

$$\alpha_f = \delta - \operatorname{atan}(v_{fy}^1/v_{fx}^1), \quad \alpha_r = -\operatorname{atan}(v_{ry}^1/v_{rx}^1)$$

with

$$\dot{v}_j^1 = R^T(\dot{x}_m^0 + \dot{R}\delta_j^1), \quad \dot{v}_i^1 = R^T(\dot{x}_m^0 + \dot{R}\delta_i^1)$$

and

$$R = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}.$$

where the superscript 1 indicates the coordinates with respect to the body fixed reference frame (\vec{e}_x^1 , \vec{e}_y^1). The dependency of the lateral tire forces on the degrees-of-freedom and the steering angle make this model quite nonlinear and difficult to analyze.

By choosing the outputs as

$$\begin{aligned} y_1 &= \dot{x}_m \cos \theta + \dot{y}_m \sin \theta - \dot{V}_d \\ y_2 &= (x_m + p \cos \theta)^2 + (y_m + p \sin \theta)^2 - R_d^2 \end{aligned}$$

with R_d the desired radius of the circle and \dot{V}_d the desired longitudinal speed, the required maneuver corresponds with a zero output. The output y_2 depends on the distance p from P to M . To solve the problem, first compute the zero dynamics and then find out its stability as a function of p .

The first problem encountered in computing the zero dynamics is the nonaffine character of the equations of motion with respect to the steering angle δ . To overcome this problem an integrator can be added to the model and δ can be regarded as a new input. Then the model has a singular decoupling matrix A , however, which makes the analysis more difficult. The `extnormform` function can be used, but this leads to insurmountable problems in the computation due to huge memory needs. Another solution is to add another integrator and use \dot{F}_{t_r} as a new input, making the decoupling matrix regular by "delaying" the input F_{t_r} also. The extension with two integrators allows the use of the standard transformation to the normal form with the function `normform`, although for a larger model. Due to the larger number of states, again problems arise with the memory needed.

To simplify the problem another "saturating" function was used in (7), $\sqrt{\cdot}$ instead of $\sin(\operatorname{atan}(\cdot))$. Now, the transformation can be computed, but the inverse transformation cannot, because of an artificial limit in the size of objects allowed by the software, putting an untimely end to the computation. Results are therefore only available in terms of the old coordinates x , but not in z .

Another attempt to solve the problem of nonaffine input is not to extend the model, but to simplify it so that the input δ appears linear in the equations. To this end, the assumption that δ is small has to be adopted. Then, by using a Taylor series expansion, the system of equations can become linear in δ if the dependency on δ of the front lateral tire force is dropped, so δ should also be small compared with α_f , an unrealistic assumption. This time the transformation to the normal form can be computed, but the set of nonlinear equations, needed to compute the inverse transformation, could not be solved, not due to memory constraints but due to limitations in the `solve` function of MAPLE. Further model simplifications are, of course, possible and may remedy this.

```
* inoutlin *
the i/o linearized system dynamics: (f1, g1)
      2              3
[ x[2]+ x[1], -2x[1]x[2] - 2x[1], 0, x[1]- x[3],
      2
x[5] + x[3] + x[1]x[2] - x[2]x[3] ]
      [ 0 0 ]
      [ 0 1 ]
      [ 1 0 ]
      [ 1 0 ]
      [ x[2] 0 ]
the explicit feedback: (alpha, beta)
      3
[ x[1] - x[3], - 2 x[1] x[2] - 2 x[1]- x[1] x[3] ]
      [ 1 0 ]
      [ 0 1 ]
```

Still another possibility is to use an algorithm that does not require model with affine inputs, see [7]. This has not been implemented.

Including several attempts to completely solve this problem were vain due to

- limits on the maximal object size
- limitations of the solve function
- limits of the available physical (or virtual) memory

The first two problems should be resolved by the MAPLE developers. The last problem is not so easy to solve because the memory requirements are likely to be exponential or double exponential with respect to the problem size [9]. More efficient algorithms are necessary, but some problems will still be computationally intractable.

Numerics can be helpful, e.g., to solve sets of nonlinear (partial differential) equations. For the vehicle problem we resorted to a purely numeric approach to find out the stability of the zero dynamics. We propose to use this problem as a benchmark or yardstick for the viability of symbolic computation programs in nonlinear control.

VI. CONCLUSION AND DISCUSSION

The computation of the normal form of the zero dynamics of the solution of the input-output and the state space exact linearization problems can be automated by using symbolic computation, e.g., by using the NonLinCon package. At the moment, the computations cannot be done for complex models due to the limited capability to solve sets of nonlinear (differential) equations. Therefore the designers of control systems cannot yet routinely compute solutions for these problems using tools based on symbolic computation programs. To remedy this, we recommend extending the capabilities of symbolic computation programs for solving large and intricate sets of nonlinear (differential) equations. It is also necessary to use more efficient algorithms. Future research will therefore aim at

- devising new or modifying existing algorithms to be more efficient in space and time
- implementing the algorithms in a more efficient way, especially with regard to computer memory requirements
- investigating a merge of symbolic with numeric computation for a hybrid solution of the problems
- solving more small and larger scale problems, to further guide in the selection of pressing lines of research

To come back to the title of the note, at the moment we cannot deny nor confirm the unqualified viability of this approach. It is expected that within this decade the capability and efficiency of symbolic computation programs are enhanced and that the increasing computing power/cost ratio of computers will enable the solution of the benchmark and larger problems at reasonable costs.

ACKNOWLEDGMENT

H. van Essen implemented most of the algorithms in MAPLE. I. van de Broek supplied the vehicle simulation problem.

REFERENCES

- [1] R. Rothfuß, J. Schaffner, and M. Zeitz, *Rechnergestützte Analyse und Synthese nichtlinearer Systeme*, in *Nichtlineare Regelung: Methoden, Werkzeuge, Anwendungen*, vol. 1026 of VDI Berichte, Düsseldorf: VDI Verlag, 1993, pp. 267–291.
- [2] O. Akhrif and G. L. Blankenship, *Computer algebra algorithms for nonlinear control*, in *Advanced Computing: Concepts and Techniques*

in *Control Engineering*, M. J. Denham and A. J. Laub, Eds., Springer-Verlag, 1988, pp. 53–80.

- [3] B. de Jager, *Nonlinear control system analysis and design*, in *Artificial Intelligence, Expert Systems and Symbolic Computation*, N. Houstis and J. R. Rice, Eds., Amsterdam: North-Holland, 1988, pp. 155–164.
- [4] H. van Essen and B. de Jager, *Analysis and design of nonlinear systems with the symbolic computation system Maple*, in *European Control Conf.*, Groningen, The Netherlands, vol. 4, Jun. 1992, pp. 2081–2085.
- [5] A. Isidori, *Nonlinear Control Systems: An Introduction*, 2nd ed., Berlin: Springer-Verlag, 1989.
- [6] H. van Essen, *Symbols speak louder than numbers: Analysis and design of nonlinear control systems with the symbolic computation system MAPLE*, Rep. WFW 92/061, Eindhoven Univ. Tech., Fac. of Mechanical Engineering, June 1992.
- [7] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, New York: Springer-Verlag, 1990.
- [8] H. B. Pacejka and I. Bakker, *The magic formula tyre model*, in *Tyre Models for Vehicle Dynamics Analysis*, Proc. 1st Int. Coll. on Tyre Models for Vehicle Dynamics Analysis, Delft, The Netherlands, 1993, pp. 1–18.
- [9] B. de Jager, *Computer aided hybrid analysis and design of nonlinear control systems*, in *Proc. 5th Int. Symp. on Application of Multivariable System Techniques*, Bradford, UK, Mar. 1994, pp. 247–254.

Note on Decentralized Adaptive Controller Design

Joon Lyoo

Abstract—This note extends the decentralized adaptive controller design of Gavel and Siljak [1] to the case when the relative order of each isolated subsystem does not exceed two.

1. INTRODUCTION

Recently Gavel and Siljak [1] presented a stable adaptive decentralized design in the face of unknown interconnection strengths and under some structural conditions placed on the interconnections. As a main result, they suggested a decentralized high gain feedback scheme for model reference adaptive control (MRAC) of an uncertain interconnected system which is composed of single input single output subsystems, and in each subsystem of which either incoming interactions enter through the control channel (Class I) or outgoing interactions pass through the measurement channel (Class II). However, the design technique for the Class I has the disadvantage of disconnecting the interconnection signal from the input of the plant. And moreover, the main result is restricted to the case when the relative order of each subsystem (n^*) is equal to one (Case 1).

This note is concerned with developing a decentralized MRAC design for an uncertain interconnected system which belongs to the Class 0 and allows for the relaxed assumption that n^* is less than and equal to two. Since in case that $n^* = 2$ (Case 2), it is not possible to choose the reference model strictly positive real any longer, the Case 2 needs to be treated in a different way from the Case 1.

Manuscript received July 24, 1992; revised March 23, 1994. This work was supported in part by the Postdoctoral Program of the Korea Science and Engineering Foundation, Republic of Korea.

The author is with the Department of Electronic Engineering, Chungnam National University, Taejeon 305-764, Korea.
IEEE Log Number 9407014.

For later comparison, the main result of [1] for the Class 0 and Case 1 is summarized in the following. Each subsystem controller consists of a precompensator C_i^p and a feedback compensator C_i^f leading to the open-loop subsystem description

$$S_i: \begin{aligned} \dot{x}_i &= A_i x_i + b_i u_i + f_i(t, y) \\ y_i &= c_i^T x_i \end{aligned} \quad (1)$$

$$C_i^p: \dot{z}_{pi} = F_i z_{pi} + g_i u_i \quad (2)$$

$$C_i^f: \dot{z}_{fi} = F_i z_{fi} + g_i y_i, \quad (3)$$

for $i = 1, 2, \dots, N$,

where $x_i \in R^{n_i}$, $z_{pi} \in R^{n_{pi}}$, and $z_{fi} \in R^{n_{fi}}$ are the states of S_i , C_i^p , and C_i^f , u_i and y_i are the scalar input and the scalar output of the subsystem S_i , and $y = [y_1, y_2, \dots, y_N]^T$. Also, f_i is a vector of output-dependent disturbances affecting S_i and satisfies the conic sector bound expressed by

$$\|f_i(t, y)\| \leq \sum_{j=1}^N \xi_{ij} |y_j|, \quad (4)$$

where ξ_{ij} 's are unknown but positive constants.

The local adaptive controllers are given by

$$u_i = \theta_i^T \nu_i, \quad i = 1, 2, \dots, N, \quad (5)$$

where $\nu_i = [\tilde{y}_i, \hat{z}_{pi}^T, \hat{z}_{fi}^T, r_i]^T$ is a vector of available signals, and $\theta_i = [\hat{d}_{fi}, \hat{c}_{pi}^T, \hat{c}_{fi}^T, \hat{k}_{oi}]^T$ is a vector of adaptation gains. Also, r_i is the reference input, and $y_i = y_i - y_{mi}$ is the model output following error with y_{mi} being the model output. $\theta_i(t)$ in (5) is automatically adjusted by the adaptation law to levels that assure stability of the overall system.

$$\dot{\theta}_i = -\Gamma_i(\sigma\theta_i + \nu_i \tilde{y}_i), \quad (6)$$

where Γ_i is a symmetric positive definite weighting matrix and σ is a positive decaying constant.

II. MAIN RESULT

To handle the Case 2, the basic control structure is modified so that the strictly positive realness of the closed-loop isolated subsystem be preserved, and the following local adaptive controllers are devised based on the design concept of [2].

$$u_i = \theta_i^T \nu_i + \hat{\theta}_i^T \Psi_{1i} - \text{sgn}(\tilde{y}_i)(\rho_i |\tilde{y}_i| + \dot{\rho}_i \Psi_{2i}) \quad (7)$$

$$\begin{aligned} \dot{\theta}_i &= -\Gamma_i(\sigma\theta_i + \Psi_{1i} \tilde{y}_i) \\ \dot{\rho}_i &= -\gamma_i(\sigma\rho_i - \Psi_{2i} |\tilde{y}_i|) \end{aligned} \quad (8)$$

$$\begin{aligned} \dot{\Psi}_{1i} &= -h_i \Psi_{1i} + \nu_i \\ \dot{\Psi}_{2i} &= -h_i \Psi_{2i} + |\tilde{y}_i| \end{aligned} \quad (9)$$

where h_i is a positive constant, γ_i is a positive weighting factor, and $\text{sgn}(\cdot)$ means the signum function. In (7), $\rho_i(t)$ is introduced to override some destabilizing effects by the interconnections from other subsystems to the i th subsystem.

Applying (7) to the nonminimal representation (1), (2), and (3), the closed-loop interconnected system becomes

$$\begin{aligned} \dot{S}: \dot{x}_i &= \hat{A}_i(\theta_i^*) \hat{x}_i + \hat{b}_i(\theta_i^T \nu_i + \hat{\theta}_i^T \Psi_{1i} - \theta_i^{*T} \nu_i) \\ &\quad - \text{sgn}(\tilde{y}_i) \hat{b}_i(\rho_i |\tilde{y}_i| + \dot{\rho}_i \Psi_{2i} - \rho_i^* |\tilde{y}_i|) \\ &\quad - \text{sgn}(\tilde{y}_i) \hat{b}_i \rho_i^* |\tilde{y}_i| + \hat{b}_i k_{oi}^* r_i - \hat{b}_i d_{fi}^* y_{mi} + f_i \\ &= \hat{A}_i(\theta_i^*) \hat{x}_i + \hat{b}_{iPR} \phi_{1i}^T \Psi_{1i} - \text{sgn}(\tilde{y}_i) \hat{b}_{iPR} \phi_{2i}^T \Psi_{2i} \\ &\quad - \text{sgn}(\tilde{y}_i) \hat{b}_{iPR} \rho_i^* \Psi_{2i} + \hat{b}_i k_{oi}^* r_i - \hat{b}_i d_{fi}^* y_{mi} + \hat{f}_i \\ y_i &= \hat{c}_i^T \hat{x}_i, \quad i = 1, 2, \dots, N. \end{aligned} \quad (10)$$

where

$$\hat{x}_i = [x_i^T, z_{pi}^T, z_{fi}^T]^T,$$

$$\hat{A}_i(\theta_i^*) = \begin{bmatrix} A_i + b_i d_{fi}^{*T} c_i^T & b_i c_{pi}^{*T} & b_i c_{fi}^{*T} \\ g_i d_{fi}^{*T} c_i^T & F_i + g_i c_{pi}^{*T} & g_i c_{fi}^{*T} \\ g_i c_i^T & 0 & F_i \end{bmatrix}$$

$$\hat{b}_i = [b_i^T, g_i^T, 0]^T, \quad \hat{c}_i^T = [c_i^T, 0, 0]$$

$$\hat{f}_i = [f_i^T, 0, 0]^T, \quad \theta_i^* = [d_{fi}^*, c_{pi}^{*T}, c_{fi}^{*T}, k_{oi}^*]^T$$

$$\phi_{1i} = \theta_i - \theta_i^*, \quad \rho_i^* > 0, \quad \phi_{2i} = \rho_i - \rho_i^*, \quad (11)$$

and \hat{b}_{iPR} is a vector satisfying $\{\hat{c}_i^T(sI_i - \hat{A}_i)^{-1} \hat{b}_i\}(s + h_i) = \{\hat{c}_i^T(sI_i - \hat{A}_i)^{-1} \hat{b}_{iPR}\}$. Note that for $\phi_i = \phi_{1i}$ or ϕ_{2i} , $(s + h_i)\phi_i(s + h_i)^{-1}$ is functionally equivalent to $\{\phi_i + \phi_i(s + h_i)^{-1}\}$ [2]. In (11), θ_i^* and ρ_i^* are unknown parameter vector and unknown positive constant to be defined later for which $\theta_i(t)$ and $\rho_i(t)$ are estimated, respectively.

Corresponding to the plant (10), the reference model chosen to be stable and of minimum phase, and such that $n_i^* = 2$ can be expressed in a nonminimal representation [1]. That is, the model transfer functions $\varphi_{mi}(s)$, $i = 1, 2, \dots, N$ can be realized as

$$\begin{aligned} \hat{M}: \dot{x}_{mi} &= \hat{A}_i(\theta_i^*) \hat{x}_{mi} + \hat{b}_i k_{oi}^* r_i \\ i &= 1, 2, \dots, N. \end{aligned} \quad (12)$$

Defining the tracking error $\hat{e}_i = \hat{x}_i - \hat{x}_{mi}$, the closed-loop error equations are derived by subtracting (12) from (10)

$$\begin{aligned} \dot{S}: \dot{\hat{e}}_i &= \hat{A}_i(\theta_i^*) \hat{e}_i + \hat{b}_{iPR} \phi_{1i}^T \Psi_{1i} - \text{sgn}(\tilde{y}_i) \hat{b}_{iPR} \phi_{2i}^T \Psi_{2i} \\ &\quad - \text{sgn}(\tilde{y}_i) \hat{b}_{iPR} \rho_i^* \Psi_{2i} - \hat{b}_i d_{fi}^* y_{mi} + \hat{f}_i \\ \dot{y}_i &= \hat{c}_i^T \hat{e}_i, \quad i = 1, 2, \dots, N. \end{aligned} \quad (13)$$

where the transfer functions $\{\hat{c}_i^T(sI_i - \hat{A}_i)^{-1} \hat{b}_{iPR}\} = \varphi_{mi}(s + h_i)$ for all i , are strictly positive real. Hence, from the Kalman-Yacubovich Lemma [1], it follows that $(\hat{A}_i, \hat{b}_{iPR}, \hat{c}_i^T)$ satisfies the equations

$$\begin{aligned} \hat{A}_i^T \hat{H}_i + \hat{H}_i \hat{A}_i &= -(N+2)I_i \\ \hat{H}_i \hat{b}_{iPR} &= \hat{c}_i^T \end{aligned} \quad (14)$$

for symmetric positive definite matrix \hat{H}_i and the identity matrix I_i .

The stability of the overall error system (13), $\dot{\phi}_{1i} = \dot{\theta}_i$ and $\dot{\phi}_{2i} = \dot{\rho}_i$ in (8) is then established through the following theorem.

Theorem: The solutions $(\hat{e}_i, \phi_{1i}, \phi_{2i})$, $\{t; t_0, \hat{e}_i(t_0), \phi_{1i}(t_0), \phi_{2i}(t_0)\}$, $i = 1, 2, \dots, N$ of the system \dot{S} , are globally uniformly bounded.

Proof: Let us choose a Lyapunov function candidate as

$$V(\hat{e}, \phi_1, \phi_2) = \sum_{i=1}^N \{\hat{e}_i^T \hat{H}_i \hat{e}_i + \phi_{1i}^T \Gamma_i^{-1} \phi_{1i} + \gamma_i^{-1} (\phi_{2i})^2\}. \quad (15)$$

Evaluating the time derivative of V along (13), (8), and (14), and then taking the norm operation, we use (4) to get

$$\begin{aligned} \dot{V} &\leq \sum_{i=1}^N \left\{ -N \|\hat{e}_i\|^2 - 2 \|\hat{e}_i\|^2 - 2 \rho_i^* |\tilde{y}_i| |\Psi_{2i}| - 2 \sigma \|\phi_{1i}\|^2 \right. \\ &\quad + 2 \sigma \|\phi_{1i}\| \|\theta_i^*\| - 2 \sigma |\phi_{2i}|^2 + 2 \sigma |\phi_{2i}| \rho_i^* \\ &\quad + 2 \|\hat{e}_i\| \|\hat{H}_i \hat{b}_i\| |d_{fi}^*| |y_{mi}| \\ &\quad \left. + 2 \|\hat{e}_i\| \|\hat{H}_i\| \sum_{j=1}^N \xi_{ij} |\tilde{y}_j| + |y_{mj}| \right\}. \end{aligned} \quad (16)$$

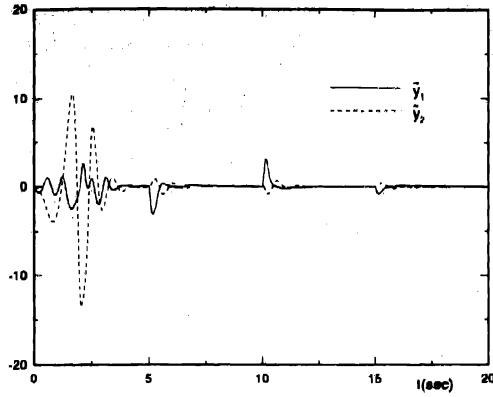


Fig. 1. Model output following errors.

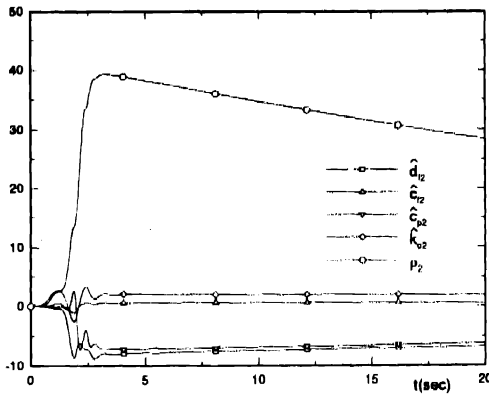


Fig. 2. Adapted gains for the subsystem 2.

Note that $\text{sgn}(\hat{y}_i) \hat{c}_i^T \hat{H}_i \hat{b}_{i,K} = \text{sgn}(\hat{y}_i) \hat{y}_i = |\hat{y}_i|$. Completing squares, (16) is modified as

$$\begin{aligned} \dot{V} \leq & \sum_{i=1}^N \left[- \left\{ \sum_{j=1}^N (\|c_i\| - \|\hat{H}_i\| \xi_{ij} |\hat{y}_j|)^2 \right\} \right. \\ & + 2\rho_i^* |\hat{y}_i|^2 - 2\rho_i^* |\hat{y}_i| \Psi_{2i} - \|\hat{c}_i\|^2 - \sigma \|\phi_{1i}\|^2 \\ & - \sigma \|\phi_{2i}\|^2 - (\|c_i\| - \mu_i)^2 + \mu_i^2 - \sigma (\|\phi_{1i}\| - \|\theta_i^*\|)^2 \\ & \left. + \sigma \|\theta_i^*\|^2 - \sigma (\|\phi_{2i}\| - \rho_i^*)^2 + \sigma (\rho_i^*)^2 \right] \end{aligned} \quad (17)$$

where ρ_i^* and μ_i^* are finite positive constants given by

$$\rho_i^* = \frac{1}{2} \sum_{j=1}^N \|\hat{H}_i\|^2 \xi_{ij}^2$$

$$\mu_i = \|\hat{H}_i \hat{b}_i\| |d_{ij}^*| \sup_t |y_{mj}| + \|\hat{H}_i\| \sum_{j=1}^N \xi_{ij} \sup_t |y_{mj}|. \quad (18)$$

In deriving ρ_i^* , $\sum_{i=1}^N \sum_{j=1}^N \|\hat{H}_i\|^2 \xi_{ij}^2 |\hat{y}_j|^2 = \sum_{j=1}^N \sum_{i=1}^N \|\hat{H}_i\|^2 \xi_{ij}^2 |\hat{y}_j|^2$ is used. The term $(-2\rho_i^* |\hat{y}_i| \Psi_{2i})$ in (17) plays the same role as the term $(-2\rho_i \hat{c}_i^T \hat{c}_i \hat{c}_i^T \hat{c}_i = 2\rho_i |\hat{y}_i|^2)$ in (5.27) of [1]. To override the destabilizing term $(2\rho_i^* |\hat{y}_i|^2)$ which arises after aggregation of all the interactions affecting the i th subsystem, Ψ_{2i} must be positive and $\Psi_{2i} \geq |\hat{y}_i|$ for all t . This is the rationale in adopting the signum and absolute functions contrary to [1].

Neglecting the negative term from (17), we obtain

$$\begin{aligned} \dot{V} \leq & \sum_{i=1}^N \{ -\|c_i\|^2 - \sigma \|\phi_{1i}\|^2 - \sigma \|\phi_{2i}\|^2 \\ & + \mu_i^2 + \sigma \|\theta_i^*\|^2 + \sigma (\rho_i^*)^2 \}. \end{aligned} \quad (19)$$

Hence, $\dot{V} < 0$ outside some compact region whose size depend on the strength of the interconnections, local design parameters, and the maximum values of the model outputs. Therefore, all signals are bounded according to Theorem 2.24 of [3].

III. NUMERICAL EXAMPLE

Consider the unstable linear constant interconnected system with unknown system parameters described by

$$\begin{aligned} \dot{x}_1 &= \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} x_1 + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_1 + \begin{bmatrix} 1 \\ -1 \end{bmatrix} y_2 \\ y_1 &= [4 \quad 2] x_1 \\ \dot{x}_2 &= \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix} x_2 + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_2 + \begin{bmatrix} 0.2 \\ -0.2 \end{bmatrix} y_2 + \begin{bmatrix} -1 \\ 1 \end{bmatrix} y_1 \\ y_2 &= [2 \quad 0] x_2. \end{aligned}$$

Note that $n_1^* = 1$ and $n_2^* = 2$. Let us choose the reference model as

$$\varphi_{m1}(s) = \frac{(s+4)}{(s+2)^2}, \quad \varphi_{m2}(s) = \frac{4}{(s+2)^2},$$

where reference inputs are square waves of height 5 and period 10 s.

Now, using the proposed decentralized adaptive scheme, computer simulations are carried out for MRAC of the example plant. The results with design parameters in (2), (3), (6), (8), and (9)

$$F_i = -4, \quad g_i = 1, \quad \Gamma_i = I_i, \quad i = 1, 2$$

$$\gamma_2 = 2, \quad h_2 = 4, \quad \sigma = 0.01,$$

are presented in Figs. 1 and 2. Due to lack of space, adapted gains are displayed only for the subsystem 2. As can be seen in the figures, the output of each subsystem closely tracks the model output after a finite interval and all the adjustable parameters are bounded for all time.

IV. CONCLUSION

Decentralized adaptive controllers have been developed for a class of interconnected systems with unknown interconnection strengths as well as uncertainties in subsystem dynamics. The local controllers essentially have output error feedback terms to override destabilizing effects by the output-dependent interactions. Specifically when the relative order of each subsystem is equal to two, each subsystem controller is reinforced such that the transfer function of the closed-loop isolated subsystem be strictly positive real.

REFERENCES

- [1] D. T. Gavel and D. D. Siljak, "Decentralized adaptive control: Structural conditions for stability," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 413-426, Apr. 1989.
- [2] K. S. Narendra and L. S. Valavani, "Stable adaptive controller design-direct control," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 570-582, Aug. 1978.
- [3] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

Characterization of Zeros in Two-Frequency-Scale Systems

Hossein M. Oloomi and Mahmoud E. Sawan

Abstract—We study the asymptotic behavior of the zeros of a two-frequency-scale transfer function in terms of the zeros of its slow and fast transfer functions. By introducing the notion of type index, we will show that all zeros of a two-frequency-scale transfer function with the type index of either $\nu \geq 0$ or $\nu = -1$ can be completely characterized in a manner that is quite similar to the characterization of the poles. We also give a partial characterization for all other zeros in terms of the cardinality of their zero sets.

I. INTRODUCTION

Single parameter standard singularity perturbed systems have been traditionally studied in the time domain [1]. However, an equivalent frequency domain characterization now exists which allows these systems to be studied entirely in the frequency domain [2]–[3]. Using this characterization, many important frequency domain problems including the two-frequency-scale (TFS) compensation design [4], the model matching problem [5]–[6], the H_∞ control theory [7]–[8], and the Hankel norm approximation problem [9] have been solved recently.

Study of the zeros of TFS systems, or equivalently the zeros of singularly perturbed systems, has not been vigorously pursued in the past. The obvious reason for this is that most often a property pertaining to singularly perturbed systems is ascertained in terms of ascertaining the same property for two lower order subsystems. However, when the property is taken to be the asymptotic behavior of the zeros, this deductive technique of proof seems to break down. The discouraging results in the following example demonstrate this point. However, despite the hopeless appearance of the results, we will show that much useful information can be obtained regarding the zeros of $H(s, \epsilon)$ from the knowledge of the zeros of $H_s(s)$ and $H_F(\epsilon s)$.

Example 1: Consider

$$H(s, \epsilon) = \frac{\epsilon s^2 + 1}{(s+1)(\epsilon s+1)}.$$

This is an example of a TFS transfer function with the slow and fast transfer functions $H_s(s) = 1/(s+1)$ and $H_F(\epsilon s) = \epsilon s/(\epsilon s+1)$, respectively. Zeros of $H(s, \epsilon)$ are at $\pm j/\sqrt{\epsilon}$. However, the slow transfer function has a zero at infinity and the fast transfer function has a zero at the origin. In contrast to what one can say about the poles, these zeros do not seem to approximate zeros of $H(s, \epsilon)$ in any meaningful sense.

From the computational point of view, until now the Newton polygon test has been the only available tool for the study of the zeros of TFS transfer functions [10]–[11]. However, this test does not take advantage of the zeros of the slow and fast transfer functions. In contrast, in this paper the behavior of the zeros of $H(s, \epsilon)$ are studied in terms of the behavior of the zeros of $H_s(s)$ and $H_F(\epsilon s)$. Since these two transfer functions have fewer zeros than $H(s, \epsilon)$, one always gains a computational advantage by taking our approach.

Manuscript received September 24, 1993; revised March 18, 1994.

H. M. Oloomi is with the Department of Electrical Engineering, Purdue University at Fort Wayne, Fort Wayne, IN 46805 USA.

M. E. Sawan is with the Department of Electrical Engineering, Wichita State University, Wichita, KS 67208 USA.

IEEE Log Number 9407015.

Additionally, we believe that our results will shed light on the performance analysis of schemes such as TFS LQG/LTR [12] and TFS H_∞ [13]–[14] theories since performance of these schemes as well as many others are ultimately tied to the location of the system zeros in the complex plane [15].

We will use the standard notation N , Z , Q , and R to denote the set of natural, integer, rational, and real numbers, respectively. The superscript “+” will be used to mean “non-negative.” For example, $Z^+ = N \cup \{0\}$. $A \setminus B$ denotes the complement of B in A , and $\text{Card}(A)$ denotes the cardinality of the finite set A . We use $\deg[\cdot]$ and $\delta[\cdot]$ to designate the degree functions for polynomials and rational functions, respectively. The set of all TFS transfer functions is denoted by T . When $H(s, \epsilon) \in T$, we denote its slow part by $H_s(s)$ and its fast part by $H_F(p)$ where $p = \epsilon s$.

We recall from Part 4 of Definition 1 in [2] that poles of a TFS transfer function are separated into two distinct groups. The first group of poles are of type $O(\epsilon^\nu)$ for some $\nu \in Q^+$ and are called the slow poles. The second group are of type $O(1/\epsilon)$ and are called the fast poles. However, as Example 1 shows, zeros of TFS systems can have more general expansions. To further explore this idea, let R_ϵ denote the ring of functions of ϵ that are analytic at $\epsilon = 0$ and let $R_s[s]$ denote the set of all polynomials in s with coefficients in $R_\epsilon[s]$. The following theorem characterizes the roots of a polynomial in $R_s[s]$ [2], [10], [16].

Theorem 1: Let $p(s, \epsilon) \in R_s[s]$. Then

- 1) Each of the n roots of $p(s, \epsilon)$ can be expanded about $\epsilon = 0$ as

$$s_i(\epsilon) = \sum_{j=1}^{\infty} b_{ij} \epsilon^{j/k}, \quad 1 \leq i \leq n, \quad (1)$$

where $b_{ij} \in C$ for each i and j , $k \in N$, $\forall \in Z$, and $\epsilon^{1/k}$ is a branch of $\epsilon^{1/k} = \epsilon$.

- 2) If one root has an expansion as in (1) with $k > 1$, then there will be $k-1$ other roots having expansions with the same coefficients but using different branches of $\epsilon^{1/k} = \epsilon$.

It is important to note that only a finite number of terms with negative exponents appear in the Laurent expansion (1). We will use Theorem 1 to classify the roots of a polynomial in $R_s[s]$ according to their Laurent expansions as given by (1).

Definition 1: A root $s_i(\epsilon)$ of $p(s, \epsilon) \in R_s[s]$ is said to be of type $O(\epsilon^\nu)$ with $\nu = N/k$, if expansion of $s_i(\epsilon)$, as given by (1), has $b_{i,1} \neq 0$ and $b_{i,j} = 0$ for $j < N$. If $b_{i,j} = 0$ for all j , then ν is taken to be zero. ν is called the type index of the root $s_i(\epsilon)$.

Note that, roots that are identically zero are classified as $O(1)$ type roots.

We will also need the concept of lost poles. Lost poles were introduced in [2] in terms of the polynomial system matrix. They were later given a state space interpretation in [3]. The following lemma is a direct consequence of Definition 2.1 in [3].

Lemma 1: Let $H(s, \epsilon) \in T$ have n_1 slow poles and n_2 fast poles. Then $H(s, \epsilon)$ will have l_1 slow lost poles iff $\delta[H_s(s)] = n_1 - l_1$, l_2 fast lost poles iff $\delta[H_F(p)] = n_2 - l_2$, and l lost poles iff it has l_1 slow lost poles and l_2 fast lost poles with $l = l_1 + l_2$.

In the next section we will study the zeros of $H(s, \epsilon) \in T$ under the following two assumptions:

- A1) $H(s, \epsilon)$ has at least one slow pole and one fast pole.
- A2) $H(s, \epsilon)$ has no lost poles.

Note that assumption (A1) excludes the single-frequency-scale case. This is no loss in generality since the single-frequency-scale

case can be treated as a special case of the TFS case. Assumption (A2), on the other hand, is not unique to this paper and in fact it is the standing assumption in the previously published works [3]–[9]. The following lemma is a direct consequence of these assumptions.

Lemma 2: Under assumptions (A1)–(A2), the transfer functions $H_S(s)$ and $H_F(p)$ are nonconstant. In particular, they are nonzero.

Proof: By (A2), $l = 0$. This implies $l_1 = l_2 = 0$. By (A1), $n_1 \geq 1$ and $n_2 \geq 1$. Hence, $\delta[H_S(s)] = n_1 - l_1 = n_1 \geq 1$ and $\delta[H_F(p)] = n_2 - l_2 = n_2 \geq 1$. Q.E.D.

II. ASYMPTOTIC BEHAVIOR OF THE ZEROS

Let $H(s, \epsilon) \in T$. Then $H(s, \epsilon)$ can be written in the fractional form

$$H(s, \epsilon) = \frac{M(s, \epsilon)}{N(s, \epsilon)}, \quad (2)$$

where $M(s, \epsilon), N(s, \epsilon) \in \mathcal{R}_\epsilon[s]$ are relatively prime. We are interested in studying the asymptotic behavior of the zeros of $H(s, \epsilon)$.

Definition 2: Define the following sets.

- The sets of zeros: Let Z_H^0 and Z_H^∞ denote the set of all zeros of $H(s, \epsilon)$ at the origin and at infinity, respectively, and let

$$Z_H := \{s \in \mathbb{C} \mid H(s, \epsilon) = 0\},$$

$$Z_H(\nu_1, \nu_2) := \{s \in Z_H \mid s = O(\epsilon^\nu), \nu_1 < \nu < \nu_2\},$$

$$Z_H(\nu) := \{s \in Z_H \mid s = O(\epsilon^\nu)\},$$

with the multiplicities included in each set.

- The sets of zero indices: Let

$$Q_Z := \{\nu \mid s \in Z_H \text{ \& } s = O(\epsilon^\nu)\},$$

$$Q_Z(\nu_1, \nu_2) := \{\nu \mid s \in Z_H \text{ \& } s = O(\epsilon^\nu), \nu_1 < \nu < \nu_2\}.$$

- The sets of poles: Let P_H^0 denote the set of all poles of $H(s, \epsilon)$ at the origin and $P_H := \{s \in \mathbb{C} \mid N(s, \epsilon) = 0\}$. Let $P_H(\nu_1, \nu_2)$ and $P_H(\nu)$ be defined similar to $Z_H(\nu_1, \nu_2)$ and $Z_H(\nu)$, respectively, but with Z_H replacing P_H . Let multiplicities be included in each of these sets.
- The sets of pole indices: Let Q_P and $Q_P(\nu_1, \nu_2)$ be defined as Q_Z and $Q_Z(\nu_1, \nu_2)$, respectively but with Z_H replacing P_H .
- For the slow and the fast transfer functions, define the set of zeros and the set of poles exactly as above but with H_S and H_F replacing H , respectively.

We are now in a position to present our main theorems. Theorem 2 shows how zeros of the slow transfer function relate to the zeros of the TFS transfer function $H(s, \epsilon)$.

Theorem 2: Let $H(s, \epsilon) \in T$ satisfy assumptions (A1)–(A2). Then

- 1) $Z_{H_S} = Z_{H_S}(0)$.
- 2) $\text{Card}(Z_{H_S}) = \sum_{\nu \in Q_Z^+} \text{Card}(Z_H(\nu))$.
- 3) $\text{Card}(Z_{H_S}^0) = \text{Card}(Z_H^0) + \text{Card}(Z_H(0, \infty))$.
- 4) Every zero $z(\epsilon) \in Z_H(\nu)$ with $\nu \in Q_Z^+$ can be approximated as $z(\epsilon) = z + r(\epsilon)$ where $z \in Z_{H_S}$ and $r(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$.

Proof: We first write $H(s, \epsilon)$ in the fractional form (2). Our immediate goal is to express both $M(s, \epsilon)$ and $N(s, \epsilon)$ in terms of their primitive factors. Since $N(s, \epsilon)$ has only slow and fast roots, its factorization is simpler. So, we proceed with $N(s, \epsilon)$ first. By definition, we have the disjoint union $Q_P = Q_P(0, \infty) \cup \{0\} \cup \{-1\}$. Therefore, we can factor $N(s, \epsilon)$ uniquely as

$$N(s, \epsilon) = g(\epsilon)N_1(s, \epsilon)N_2(s, \epsilon)N_3(s, \epsilon), \quad (3)$$

where $g(\epsilon)$ is analytic at $\epsilon = 0$, and $N_1(s, \epsilon)$, $N_2(s, \epsilon)$ and $N_3(s, \epsilon)$ are monic polynomials (in s) with roots that have type indices in the sets $Q_P(0, \infty)$, $Q_P(0)$, $Q_P(-1)$, respectively. To obtain a more useful representation for $N(s, \epsilon)$, let $\epsilon^\nu \beta_{j,\nu}(\epsilon)$ denote the j th root of $N(s, \epsilon)$ among those roots that have type index ν . Assuming the existence of such roots, we define the following two families of polynomials

$$q_\nu(s, \epsilon) := \prod_{j=1}^{l_\nu} (s - \epsilon^\nu \beta_{j,\nu}(\epsilon)), \quad \nu \in Q_P, \quad (4)$$

$$\tilde{q}_\nu(s, \epsilon) := \epsilon^{-\nu l_\nu} q_\nu(s, \epsilon). \quad (5)$$

We observe that $q_\nu(s, \epsilon) \in \mathcal{R}_\epsilon[s]$ whenever $\nu \in Q_P^+$, and $\tilde{q}_\nu(s, \epsilon) \in \mathcal{R}_\epsilon[s]$ whenever $\nu \in Q_P \setminus Q_P^+$. Use of (4)–(5) allows us to write

$$N_1(s, \epsilon) = \prod_{\nu \in Q_P(0, \infty)} q_\nu(s, \epsilon), \quad (6)$$

$$N_2(s, \epsilon) = q_0(s, \epsilon) = s^{n_{00}} \prod_{j=1}^{l_0 - n_{00}} (s - \beta_{j,0}(\epsilon)), \quad \beta_{j,0}(\epsilon) \neq 0$$

$$:= s^{n_{00}} N_{20}(s, \epsilon), \quad (7)$$

$$N_3(s, \epsilon) = q_{-1}(s, \epsilon) = \epsilon^{-l_{-1}} \tilde{q}_{-1}(s, \epsilon)$$

$$:= \epsilon^{-n_3} \tilde{N}_3(s, \epsilon), \quad n_3 := l_{-1}, \quad (8)$$

where $n_j = \deg[N_j(s, \epsilon)]$ for $j = 1, 2, 4$, and by definition $\beta_{j,-1}(0) \neq 0$ [2]. Thus, $N(s, \epsilon)$ can be factored alternatively as

$$N(s, \epsilon) = \epsilon^{-n_3} g(\epsilon) [N_1(s, \epsilon) s^{n_{00}}] N_{20}(s, \epsilon) \tilde{N}_3(s, \epsilon). \quad (9)$$

Next, we obtain a factorization for $M(s, \epsilon)$. Since each root of $M(s, \epsilon)$ has an expansion of the form (1) given by Theorem 1, we factor $M(s, \epsilon)$ as

$$M(s, \epsilon) = f(\epsilon) \prod_{j=1}^5 M_j(s, \epsilon), \quad (10)$$

where $f(\epsilon)$ is analytic at $\epsilon = 0$, and $M_j(s, \epsilon)$ for $j = 1, \dots, 5$ are monic polynomials (in s) and have roots with type indices in the sets $Q_Z(0, \infty)$, $\{0\}$, $Q_Z(-1, 0)$, $\{-1\}$, and $Q_Z(-\infty, -1)$, respectively. Since these sets are pairwise disjoint and have union equal to Q_Z , this factorization is unique. An alternative representation for $M(s, \epsilon)$ is obtained by introducing the following two families of polynomials

$$p_\nu(s, \epsilon) := \prod_{j=1}^{k_\nu} (s - \epsilon^\nu \alpha_{j,\nu}(\epsilon)), \quad \nu \in Q_Z, \quad (11)$$

$$\hat{p}_\nu(s, \epsilon) := \epsilon^{-\nu k_\nu} p_\nu(s, \epsilon), \quad (12)$$

where $\epsilon^\nu \alpha_{j,\nu}(\epsilon)$ denotes the j th root of $M(s, \epsilon)$ among those that have type index ν , and k_ν denotes the number of such roots. We observe that $p_\nu(s, \epsilon) \in \mathcal{R}_\epsilon[s]$ whenever $\nu \in Q_Z^+$, and $\hat{p}_\nu(s, \epsilon) \in \mathcal{R}_\epsilon[s]$ whenever $\nu \in Q_Z \setminus Q_Z^+$. Use of (11)–(12) allows us to write

$$M_1(s, \epsilon) = \prod_{\nu \in Q_Z(0, \infty)} p_\nu(s, \epsilon), \quad (13)$$

$$M_2(s, \epsilon) = p_0(s, \epsilon) = s^{m_{00}} \prod_{j=1}^{k_0 - m_{00}} (s - \alpha_{j,0}(\epsilon)), \quad \alpha_{j,0}(\epsilon) \neq 0$$

$$:= s^{m_{00}} M_{20}(s, \epsilon), \quad (14)$$

$$M_3(s, \epsilon) = \prod_{\nu \in Q_Z(-1, 0)} p_\nu(s, \epsilon) = \prod_{\nu \in Q_Z(-1, 0)} \epsilon^{\nu k_\nu} \hat{p}_\nu(s, \epsilon)$$

$$:= \epsilon^{-m_3} \prod_{\nu \in Q_Z(-1, 0)} \hat{p}_\nu(s, \epsilon) =: \epsilon^{-m_3} \tilde{M}_3(s, \epsilon), \quad (15)$$

$$M_4(s, \epsilon) = p_{-1}(s, \epsilon) = e^{-k-1} \tilde{p}_{-1}(s, \epsilon) = \epsilon^{-m_4} \tilde{M}_4(s, \epsilon),$$

$$m_4 = k_{-1}, \quad (16)$$

$$M_5(s, \epsilon) = \prod_{\nu \in Q_Z(-\infty, -1)} p_\nu(s, \epsilon) = \prod_{\nu \in Q_Z(-\infty, -1)} \epsilon^{\nu k_\nu} \tilde{p}_\nu(s, \epsilon)$$

$$:= \epsilon^{-\mu_5} \prod_{\nu \in Q_Z(-\infty, -1)} \tilde{p}_\nu(s, \epsilon) = \epsilon^{-\mu_5} \tilde{M}_5(s, \epsilon), \quad (17)$$

where $m_j = \deg[M_j(s, \epsilon)]$ for $j = 1, \dots, 5$, and $\mu_3, \mu_5 \in \mathbb{N}$ by Part 2 of Theorem 1. Hence, $M(s, \epsilon)$ can be factored alternatively as

$$M(s, \epsilon) = \epsilon^{-(\mu_3+m_4+\mu_5)} f(\epsilon) [M_1(s, \epsilon) s^{m_{00}}] \cdot M_{20}(s, \epsilon) \tilde{M}_1(s, \epsilon) \tilde{M}_4(s, \epsilon) \tilde{M}_5(s, \epsilon), \quad (18)$$

and using (9) and (18) we can express $H(s, \epsilon)$ as

$$H(s, \epsilon) = \epsilon^{\gamma_5} \frac{f(\epsilon) M_1(s, \epsilon) s^{m_{00}} M_{20}(s, \epsilon) \tilde{M}_1(s, \epsilon) \tilde{M}_4(s, \epsilon) \tilde{M}_5(s, \epsilon)}{g(\epsilon) N_1(s, \epsilon) s^{n_{00}} N_{20}(s, \epsilon) \tilde{N}_1(s, \epsilon)}, \quad (19)$$

where

$$\gamma_5 = n_4 - (\mu_3 + m_4 + \mu_5). \quad (20)$$

We are now in a position to analyze the limit of $H(s, \epsilon)$ as $\epsilon \rightarrow 0$. We observe that as ϵ approaches zero, each of the polynomials $\tilde{M}_1(s, \epsilon)$, $\tilde{M}_4(s, \epsilon)$, and $\tilde{M}_5(s, \epsilon)$ approaches a finite constant limit, say K_j , $j=3, 4, 5$, respectively. Since $H(s, \epsilon)$ has no lost poles, these constants are nonzero according to Lemma 2. Since $H(s, 0)$ is defined, the polynomial $\tilde{N}_4(s, \epsilon)$ must also approach a finite nonzero limit, say K' . Moreover, $M_1(s, \epsilon) \rightarrow s^{m_1}$, $M_{20}(s, \epsilon) \rightarrow M_{20}(s, 0) = \prod_{j=1}^{m_2-m_{00}} (s - \alpha_{j,0}(0))$ where $\alpha_{j,0}(0) \neq 0$, $N_1(s, \epsilon) \rightarrow s^{n_1}$, and $N_{20}(s, \epsilon) \rightarrow N_{20}(s, 0) = \prod_{j=1}^{n_2-n_{00}} (s - \beta_{j,0}(0))$ where $\beta_{j,0}(0) \neq 0$. Therefore,

$$\frac{M_1(s, \epsilon) s^{m_{00}}}{N_1(s, \epsilon) s^{n_{00}}} \rightarrow s^{\lambda_5}$$

where,

$$\lambda_5 = (m_{00} + m_1) - (n_{00} + n_1). \quad (21)$$

Finally, since $H(s, 0)$ is defined, the quantity $\epsilon^{\gamma_5} f(\epsilon)/g(\epsilon)$ must approach a finite limit, say $K/K'/\prod_{j=1}^5 K_j$, and in the light of Lemma 2, this limit must be nonzero. Hence,

$$H_5(s) := H(s, 0) = K_5 \frac{s^{\lambda_5} M_{20}(s, 0)}{N_{20}(s, 0)}. \quad (22)$$

Since all zeros of $H_5(s)$ are $O(1)$ type zeros, Part 1 of the theorem follows immediately. To prove Part 3, first note from (21) that although the quantities $(m_{00} + m_1)$ and $(n_{00} + n_1)$ are both positive, they cannot be simultaneously strictly positive since otherwise, there would be at least one pole-zero cancellation at the origin and this contradicts (A2). Using this fact we observe that $\lambda_5 < 0$ iff $(n_{00} + n_1) > 0$ and $(m_{00} + m_1) = 0$. Hence, $\text{Card}(Z_{H_5}^0) = m_{00} + m_1 = 0$ and the result holds trivially in this case. On the other hand, $\lambda_5 \geq 0$ iff $(m_{00} + m_1) \geq 0$ and $(n_{00} + n_1) = 0$, and in this case

$$\begin{aligned} \text{Card}(Z_{H_5}^0) &= m_{00} + m_1 \\ &= m_{00} + \deg[M_1(s, \epsilon)] \\ &= \text{Card}(Z_{H_1}^0) + \text{Card}(Z_H(0, \infty)). \end{aligned}$$

This completes the proof of Part 3. To prove Part 2, we note that

$$\begin{aligned} \text{Card}(Z_{H_5}) &= \text{Card}(Z_{H_5}^0) + \deg[M_{20}(s, 0)] \\ &= (m_{00} + m_1) + (m_2 - m_{00}) \\ &= m_1 + m_2 \\ &= \sum_{\nu \in Q_Z^+} \text{Card}(Z_H(\nu)). \end{aligned}$$

Finally, suppose $z(\epsilon) \in Z_H(\nu)$ for some $\nu \in Q_Z^+$. Then $z(\epsilon)$ is a root of the polynomial $M_1(s, \epsilon) M_2(s, \epsilon)$. Since $z(\epsilon)$ has type index $\nu \geq 0$, $z(\epsilon)$ is analytic at $\epsilon = 0$. Hence, $z(\epsilon) = z(0) + r(\epsilon)$ where $r(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$. Our foregoing analysis shows that $M_1(s, \epsilon) M_2(s, \epsilon)$ approaches $s^{m_1+m_{00}} M_{20}(s, 0)$, which is the zero polynomial of $H_5(s)$. Hence, $z := z(0) \in H_5(s)$. This proves the last part of the theorem. Q.E.D.

The next theorem relates the zeros of the fast transfer function to the zeros of the TFS transfer function $H(s, \epsilon)$.

Theorem 3: Let $H(s, \epsilon) \in T$ satisfy assumptions (A1)–(A2). Then

- 1) $Z_{H_5} = Z_{H_5}^0 \cup Z_{H_5}(-1)$.
- 2) $\text{Card}(Z_{H_5}) = \sum_{\nu \in Q_Z(-1, \infty) \cup \{-1\}} \text{Card}(Z_H(\nu)) - \sum_{\nu \in Q_Z(-1, \infty)} \text{Card}(P_H(\nu))$.
- 3) $\text{Card}(Z_{H_5}^0) = \sum_{\nu \in Q_Z(-1, \infty)} \text{Card}(Z_H(\nu)) - \sum_{\nu \in Q_Z(-1, \infty)} \text{Card}(P_H(\nu))$.
- 4) Every zero $z(\epsilon) \in Z_H(-1)$ can be approximated as $z(\epsilon) = z + r(\epsilon)$ where $z \in Z_{H_5}(-1)$ and $r(\epsilon) \rightarrow 0$.

Proof: Due to the space limitation, we only outline the proof. Using the notation and results from the previous theorem, we first express $H(p/\epsilon, \epsilon)$ as shown by (23), found at the bottom of the page, where

$$\gamma_4 = (n_1 + n_2 + n_4) - (m_1 + m_2 + m_3 + m_4 + \mu_5). \quad (24)$$

By the repeated use of Part 2 of Theorem 1, we observe that as $\epsilon \rightarrow 0$, $\epsilon^{m_1} M_j(p/\epsilon, \epsilon) \rightarrow p^{m_j}$ and $\epsilon^{n_j} N_j(p/\epsilon, \epsilon) \rightarrow p^{n_j}$ for $j=1, 2$, $\epsilon^{m_3-\mu_3} \tilde{M}_3(p/\epsilon, \epsilon) \rightarrow p^{m_3}$, $\tilde{M}_4(p/\epsilon, \epsilon) \rightarrow \prod_{j=1}^{m_4} (p - \alpha_{j-1}(0))$ where $\alpha_{j-1}(0) \neq 0$, $\tilde{M}_5(p/\epsilon, \epsilon) \rightarrow K_5$, and $\tilde{N}_4(p/\epsilon, \epsilon) \rightarrow \prod_{j=1}^{n_4} (p - \beta_{j-1}(0))$ where $\beta_{j-1}(0) \neq 0$. Thus, taking the limit in (23) as $\epsilon \rightarrow 0$, we obtain

$$\begin{aligned} H_I(p) &:= H\left(\frac{p}{\epsilon}, \epsilon\right) \Big|_{\epsilon=0} \\ &= \epsilon^{\gamma_4} \frac{f(\epsilon)}{g(\epsilon)} \Big|_{\epsilon=0} \frac{p^{\lambda_4} \tilde{M}_4\left(\frac{p}{\epsilon}, \epsilon\right) \Big|_{\epsilon=0}}{\tilde{N}_4\left(\frac{p}{\epsilon}, \epsilon\right) \Big|_{\epsilon=0}} K_5, \end{aligned} \quad (25)$$

where,

$$\lambda_4 = (m_1 + m_2 + m_3) - (n_1 + n_2). \quad (26)$$

Since $H(s, \epsilon)$ has no lost poles, $K_5 \neq 0$ by Lemma 2. Additionally, since $H_I(p)$ is defined, $\epsilon^{\gamma_4} f(\epsilon)/g(\epsilon)$ must approach a finite nonzero constant, say K_I/K_5 . Hence, we can represent $H_I(p)$ as

$$H_I(p) = K_I \frac{p^{\lambda_4} \tilde{M}_4\left(\frac{p}{\epsilon}, \epsilon\right) \Big|_{\epsilon=0}}{\tilde{N}_4\left(\frac{p}{\epsilon}, \epsilon\right) \Big|_{\epsilon=0}}. \quad (27)$$

$$H\left(\frac{p}{\epsilon}, \epsilon\right) = \epsilon^{\gamma_4} \frac{f(\epsilon)}{g(\epsilon)} \frac{[\epsilon^{m_1} M_1\left(\frac{p}{\epsilon}, \epsilon\right)] [\epsilon^{m_2} M_2\left(\frac{p}{\epsilon}, \epsilon\right)] [\epsilon^{m_3-\mu_3} \tilde{M}_3\left(\frac{p}{\epsilon}, \epsilon\right)] \tilde{M}_4\left(\frac{p}{\epsilon}, \epsilon\right) \tilde{M}_5\left(\frac{p}{\epsilon}, \epsilon\right)}{[\epsilon^{n_1} N_1\left(\frac{p}{\epsilon}, \epsilon\right)] [\epsilon^{n_2} N_2\left(\frac{p}{\epsilon}, \epsilon\right)] \tilde{N}_4\left(\frac{p}{\epsilon}, \epsilon\right)} \quad (23)$$

To prove Part 1 of the theorem, we must first show that $\lambda_L \geq 0$. If not, suppose to the contrary $\lambda_L < 0$. Since $H(s, \epsilon) \in T_c$, by definition $H_I(p)$ cannot have a pole at the origin. Therefore, see from (27) that $M_4(z, \epsilon)|_{z=0}$ must have λ_L roots at the origin to cancel these poles. But this leads to a contradiction since $\lambda_L(0) \neq 0$. Hence, $\lambda_L \geq 0$ i.e. all zeros of $H_I(s, \epsilon)$ are either at the origin or have type the index $\nu = -1$. To prove Part 2, we see that

$$\begin{aligned} \sum_{\epsilon \in Q_H(-1) \cup \{1\}} \text{Card}(Z_H(\nu)) &= \sum_{\epsilon \in Q_H(-1) \cup \{1\}} \text{Card}(P_H(\nu)) \\ &= \sum_1 m_1 - \sum_1 n_1 \\ &= \lambda_L + m_1 \\ &= \lambda_L + \deg M_1(p/\epsilon) \\ &= \text{Card}(Z_{H_I}) \end{aligned}$$

Part 3 follows easily by subtracting $m_1 = \text{Card}(Z_H(-1))$ from both sides of the above equation. Finally let $\epsilon(\epsilon) \in Z_H(-1)$. Then $\epsilon(\epsilon)$ is a root of $M_1(s, \epsilon)$. Since $\epsilon(\epsilon)$ has the type index $\nu = -1$, $\epsilon(\epsilon)$ is analytic at $\epsilon = 0$ and we can write $\epsilon(\epsilon) = \epsilon_1 + \epsilon_2(\epsilon)$ where ϵ_1 is a constant and $\epsilon_2(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$. Part 4 now follows by setting $\epsilon(\epsilon) = \epsilon_2(\epsilon)/\epsilon = \epsilon_2/\epsilon$ and noting that ϵ_1 is a root of $M_1(p/\epsilon) |_{\epsilon=0}$. Q.E.D.

By studying Part 4 of Theorems 2 and 3 we see that the asymptotic behavior of the two polynomials $M_1(s, \epsilon)$, $M(s, \epsilon)$ and $M_4(s, \epsilon)$ are completely characterized in terms of the zeros of the slow and the fast transfer functions respectively. However to completely characterize the asymptotic behavior of all zeros of $H(s, \epsilon)$ the asymptotic behavior of the two polynomials $M_1(s, \epsilon)$ and $M(s, \epsilon)$ in (18) should also be studied in terms of the behavior of the zeros of the slow and fast transfer functions. Although a complete characterization cannot be given due to the vanishing behavior of these two polynomials some partial information can still be obtained as the following two theorems show. To state the first theorem note that $M_1(s, \epsilon)$ in (15) can be written as

$$M_1(s, \epsilon) = \epsilon^l s^{n_1} + \epsilon M(s, \epsilon) \quad \epsilon_1(\epsilon) = M_1(0, \epsilon) \quad (28)$$

where $M(s, \epsilon) \in R[s]$ has degree less than m_1 and $\epsilon_1(0) \neq 0$. Theorem 4 shows that m_1 , μ_1 and $\epsilon_1(0)$ can be determined from the knowledge of the slow and fast transfer functions.

Theorem 4 Under assumptions (A1)–(A2)

- 1) $\mu_1 = \text{Card}(Z_{H_I}^0)$
- 2) $m_1 = \text{Card}(Z_H(-1, 0)) = \text{Card}(Z_{H_I}^*) + \text{Card}(Z_{H_I}^0)$
- 3) If $m_1 \neq 0$ $\epsilon_1(0) = \frac{\lim_{\epsilon \rightarrow 0} \prod_{i=1}^{m_1} \epsilon_i(\epsilon)}{\lim_{\epsilon \rightarrow 0} \prod_{i=1}^{m_1} \epsilon_i(\epsilon)}$ Otherwise $\epsilon_1(0) = \frac{\lim_{\epsilon \rightarrow 0} \prod_{i=1}^{m_1} \epsilon_i(\epsilon)}{\lim_{\epsilon \rightarrow 0} \prod_{i=1}^{m_1} \epsilon_i(\epsilon)}$

Proof In the course of proving Theorems 2 and 3 we proved that $\epsilon = f(\epsilon)/g(\epsilon)$ and $\epsilon^{-1}f(\epsilon)/g(\epsilon)$ approach some finite nonzero constants. This implies that $\epsilon = \epsilon_1$. Thus by equating (20) and (24), we obtain

$$\mu_1 = m_1 + m_2 + m_3 - (n_1 + n_2) \quad (29)$$

On the other hand Part 3 of Theorem 3 shows that

$$\text{Card}(Z_{H_I}^0) = m_1 + m_2 + m_3 - (n_1 + n_2) \quad (30)$$

Comparing (29) and (30) proves Part 1 of the theorem. For Part 2 we rearrange (29) as $m_1 = [(n_1 + n_2) - (m_1 + m_2)] + \mu_1$ and identify the term in the bracket with $\text{Card}(Z_{H_I}^*)$. The last part follows from (22) and (27) noting that $\lambda_L = \epsilon_1(0)$. Q.E.D.

The next theorem shows that $m^* = \deg[M(s, \epsilon)]_{\epsilon=0}$ is determined. Noting that $m^* = \text{Card}(Z_H(-\infty, -1))$ we have

Theorem 5 Let $H(s, \epsilon)$ satisfy assumptions (A1)–(A2). Then

$$\text{Card}(Z_H(-\infty, -1)) = \text{Card}(Z_{H_I}^*) - \text{Card}(H)$$

Proof We have

$$\begin{aligned} \text{Card}(Z_H) &= \sum_{\epsilon \in Q_H} \text{Card}(Z_H(\nu)) \\ &= \sum_{\epsilon \in Q_H^*} \text{Card}(Z_H(\nu)) \\ &\quad + \left[\sum_{\epsilon \in Q_H(-1) \cup \{1\}} \text{Card}(Z_H(\nu)) \right. \\ &\quad \left. - \sum_{\epsilon \in Q_H(-1) \cup \{1\}} \text{Card}(Z_H(\nu)) \right] \\ &\quad + \sum_{\epsilon \in Q_H(-1, 0)} \text{Card}(Z_H(\nu)) \\ &\quad + \sum_{\epsilon \in Q_H(-\infty, -1)} \text{Card}(Z_H(\nu)) \end{aligned}$$

We now replace the first and the fourth summations with the quantities given by Part 2 of Theorem 2 and Part 2 of Theorem 4 respectively. We also replace the summation in the bracket with the quantity obtained after subtracting Part 3 from Part 2 in Theorem 3. After simplifying the result we obtain

$$\begin{aligned} \text{Card}(Z_H) &= \text{Card}(Z_{H_I}^*) + \text{Card}(Z_{H_I}^0) + \text{Card}(Z_{H_I}^*) \\ &\quad + \text{Card}(Z_H(-\infty, -1)) \end{aligned}$$

A further simplification is made upon substituting $\text{Card}(P_H) = \text{Card}(Z_H)$ for the quantity $\text{Card}(Z_{H_I}^0)$ and noting that $\text{Card}(P_H) = \text{Card}(Z_H) - \text{Card}(Z_{H_I}^*)$. This yields

$$\begin{aligned} \text{Card}(Z_H(-\infty, -1)) &= [\text{Card}(P_H) - \text{Card}(Z_{H_I}^*)] \\ &\quad - [\text{Card}(P_H) - \text{Card}(Z_H)] \end{aligned}$$

The theorem now follows by identifying the terms in the brackets with $\text{Card}(Z_{H_I}^*)$ and $\text{Card}(Z_{H_I}^0)$ respectively. Q.E.D.

It is important to note that in determining the exact value of m the relative degree of $H(s, \epsilon)$ should be known in general. In other words knowing only the fast transfer function does not help one to determine m in general as stated by Theorem 5. However if the relative degree of $H(s, \epsilon)$ is not known *a priori* an upper bound for m can still be obtained as the following corollary shows.

Corollary 1 Let $H(s, \epsilon)$ satisfy assumptions (A1)–(A2). Then

$$\text{Card}(Z_H(-\infty, -1)) \leq \text{Card}(Z_{H_I}^*)$$

There is of course a special case where the knowledge of the fast transfer function is sufficient to determine m . This special case is treated in the following corollary.

Corollary 2 Let $H(s, \epsilon)$ satisfy assumptions (A1)–(A2). If $H_I(p)$ is not strictly proper, then

- 1) $\text{Card}(Z_{H_I}^*) = 0$ i.e. $H(s, \epsilon)$ is not strictly proper either
- 2) $\text{Card}(Z_H(-\infty, -1)) = 0$

Proof Since $H_I(p)$ is proper but not strictly proper, $\text{Card}(Z_{H_I}^*) = 0$. By Theorem 5 this implies $\text{Card}(Z_H(-\infty, -1)) = -\text{Card}(Z_{H_I}^*)$. The results now follow. Q.E.D.

Theorem 5 can also be used to prove a scalar version of Lemma 2.5 in [4].

Corollary 3 Let $H(s, \epsilon)$ satisfy assumptions (A1)–(A2). Then $H^{-1}(s, \epsilon) \in T$ iff $H(s, \infty) \neq 0$ and $H_I(\infty) \neq 0$.

Proof: (\Rightarrow): Since $H_s(s)$ and $(H^{-1}(s, \epsilon))_s = (H_s(s))^{-1}$ are proper, $\text{Card}(Z_{H_s}^\infty) = 0$. On the other hand, properness of $H(s, \epsilon)$ and $H^{-1}(s, \epsilon)$ imply that $\text{Card}(Z_H^\infty) = 0$. Since $H^{-1}(s, \epsilon) \in T_+$, we also have $\text{Card}(Z_H(-\infty, -1)) = 0$ and $\text{Card}(Z_H(-1, 0)) = 0$. Hence, by Theorem 5, $\text{Card}(Z_{H_F}^\infty) = 0$.

(\Leftarrow): Since $H_F(\infty) \neq 0$, Part 1 of Corollary 2 shows that $\text{Card}(Z_H^\infty) = 0$. Hence, $H^{-1}(s, \epsilon)$ is proper. Part 2 of Corollary 2 shows that $\text{Card}(Z_H(-\infty, -1)) = 0$. Hence, for $H^{-1}(s, \epsilon)$ to be in T_+ , it suffices to show that $\text{Card}(Z_H(-1, 0)) = 0$. But this follows since $H_F(0) = H_s(\infty) \neq 0$ implies $\text{Card}(Z_{H_F}^0) = 0$ and $\text{Card}(Z_{H_F}^\infty) = 0$. So, by Theorem 4, $\mu_3 = m_3 = 0$. Q.E.D.

Example 2: Reconsider Example 1. Suppose that $H_s(s)$ and $H_F(p)$ are given. Suppose it is also known that $H(s, \epsilon)$ has no lost poles. Since the slow transfer function has no finite zeros, Parts 1 and 4 of Theorem 2 are vacuously true. Part 2 of that theorem shows that $H(s, \epsilon)$ has no zeros with type index $\nu \geq 0$. In particular, $H(s, \epsilon)$ has no zeros at the origin. Since $Z_{H_F}(-1) = 0$, Part 4 of Theorem 3 is vacuously true. Part 2 of Theorem 3 shows that $H(s, \epsilon)$ must have two zeros with type indices in the interval $(-1, \infty)$. Since $H(s, \epsilon)$ has no zeros with type index $\nu \geq 0$, we conclude that $H(s, \epsilon)$ must have two zeros with type indices in the interval $(-1, 0)$. This result can also be verified from Theorem 4 which gives $m_1 = 1$, $\nu_1 = 2$, and $\dot{\nu}_3(0) = 1$. Moreover, Corollary 2 applies to the present case. Therefore, we conclude that $H(s, \epsilon)$ has a zero polynomial of the form $\epsilon s^2 + r_M(s, \epsilon)$ where $r_M(s, \epsilon) \in \mathcal{R}_r[s]$ has degree of either zero or one and $r_M(0, 0) = 1$.

Remark 1: The results obtained in this paper are also useful when studying the asymptotic behavior of the poles of a certain class of singularly perturbed system which are more general than TFS systems [17]. To see this, let $H(s, \epsilon)$ be the transfer function of a (not necessarily standard) singularly perturbed system. Assume that $H^{-1}(s, \epsilon) \in T_+$. In particular, this means that $H(s, \epsilon)$ is invertible. From [3], we know that $H^{-1}(s, \epsilon)$ has a standard singularly perturbed state space realization

$$\left(\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, [C_1 \ C_2], D \right),$$

where all matrices are functions of ϵ which are analytic at $\epsilon = 0$ and $A_{22}(0)$ is invertible. The original system $H(s, \epsilon)$ has the realization

$$\left(\begin{bmatrix} A_{11}^x & A_{12}^x \\ A_{21}^x & A_{22}^x \end{bmatrix}, \begin{bmatrix} B_1 D^{-1} \\ B_2 D^{-1} \end{bmatrix}, [-D^{-1} C_1 \ -D^{-1} C_2], D^{-1} \right),$$

where $A_{ij}^x := A_{ij} - B_i D^{-1} C_j$ for $i, j = 1, 2$. Hence, when $A_{22}^x(0)$ is invertible, the original system is a TFS system and its poles have the usual TFS behavior. However, when $A_{22}^x(0)$ is not invertible, one gets more complicated asymptotic behavior as described in this paper. Since poles of the original system are zeros of the inverse system, the asymptotic behavior of poles of $H(s, \epsilon)$ can be studied in terms of the asymptotic behavior of the zeros of $H^{-1}(s, \epsilon)$ using the results obtained in this paper.

REFERENCES

- [1] P. Kokotovic, K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*. Orlando: Academic Press, 1986.
- [2] D. W. Luse and H. K. Khalil, "Frequency domain results for systems with slow and fast dynamics," *IEEE Trans. Automat. Contr.*, vol. AC-30, no. 12, pp. 1171-1179, 1985.
- [3] D. W. Luse, "State-space realization of multiple-frequency-scale transfer matrices," *IEEE Trans. Automat. Contr.*, vol. AC-33, no. 2, pp. 185-187, 1988.

- [4] H. K. Khalil, "Output feedback control of linear two-time-scale systems," *IEEE Trans. Automat. Contr.*, vol. AC-32, no. 9, pp. 784-792, 1987.
- [5] H. M. Oloomi and M. E. Sawan, "An algorithm for computing the two-frequency-scale model matching compensator," *Proc. Amer. Contr. Conf.*, pp. 1640-1641, 1991.
- [6] —, "Suboptimal model-matching problem for two frequency scale transfer functions," *Proc. Amer. Contr. Conf.*, pp. 2190-2191, 1989.
- [7] D. W. Luse and J. A. Ball, "Frequency-scale-decomposition for H^∞ -disk problems," *SIAM J. Contr. Optim.*, vol. 27, pp. 814-835, 1989.
- [8] Z. Pan and T. Basar, " H^∞ -optimal control for singularly perturbed systems, part I: Perfect state measurements," *Proc. Amer. Contr. Conf.*, pp. 1850-1854, 1992.
- [9] H. M. Oloomi and M. E. Sawan, "Spectral approximation of almost block triangular operators and Hankel norm approximation of two frequency scale transfer matrices," *Proc. Annual Allerton Conf.*, pp. 903-904, 1988.
- [10] D. K. Knopp, *Theory of Functions, part II*. New York: Dover, 1947.
- [11] G. A. Bliss, *Algebraic Functions*. New York: AMS Colloquium, vol. VI, 1993.
- [12] G. Stein and M. Athans, "The LQG/LTR procedure for multivariable feedback control design," *IEEE Trans. Automat. Contr.*, vol. AC-32, no. 2, pp. 105-114, 1987.
- [13] B. M. Chen, A. Saberi, P. Sannuti, and Y. Shamash, "Construction and parametrization of all static and dynamic H_2 -optimal state feedback solutions, optimal fixed modes, and fixed decoupling zeros," *IEEE Trans. Automat. Contr.*, vol. 38, no. 2, pp. 248-261, 1993.
- [14] D. J. N. Limebeer and Y. S. Hung, "An analysis of the pole-zero cancellation in H^∞ optimal control problem of the first kind," *SIAM J. Contr. Optim.*, vol. 25, pp. 1457-1493, 1987.
- [15] A. G. MacFarland, *Complex Variable Methods for Linear Multivariable Feedback Systems*. London: Taylor and Francis, 1980.
- [16] E. Hille, *Analytic Function Theory*. Ginn and Company, 1962.
- [17] H. K. Khalil, "Feedback control of nonstandard singularly perturbed systems," *IEEE Trans. Automat. Contr.*, vol. AC-34, no. 10, pp. 1052-1060, 1989.

Semi-Global Stabilizability of Linear Null Controllable Systems with Input Nonlinearities

Andrew R. Teel

Abstract—An H_∞ -based Lyapunov proof is provided for a result recently established by Lin and Saberi: if a linear system is asymptotically null controllable with bounded controls then, when subject to input saturation, it is semi-globally stabilizable by linear state feedback. A new result is that if the system is also detectable then it is semi-global stabilizable by completely linear output feedback. Further, an extension which relaxes the requirements on the input characteristic is obtained.

1. INTRODUCTION

There has been much interest in the last few years concerning stabilization of linear systems subject to actuator saturation. Several important papers have focused on systems which are null controllable with bounded controls. A linear system is said to be asymptotically null controllable with bounded controls if any initial condition can be

Manuscript received September 27, 1993; revised March 30, 1994. Research supported in part by research funds of the Graduate School of the University of Minnesota and by NSF Grant ECS-9309523.

The author is with the Department of Electrical Engineering, University of Minnesota, 4-174 EE/CS Building, 200 Union St., SE, Minneapolis, MN 55455 USA.

IEEE Log Number 9407016.

driven to the origin, perhaps in the limit as $t \rightarrow \infty$, using arbitrarily small controls. It is well known that asymptotic null controllability with bounded controls is equivalent to the usual notion of linear stabilizability plus the added condition that the open loop eigenvalues have nonpositive real part. For a thorough treatment of this concept see [7] or [8]. It is also well known that, in general, a linear feedback cannot be used to globally asymptotically stabilize the origin of a linear system null controllable with bounded controls if the system has input constraints. This was first pointed out in [3] and elaborated on in [10]. On the other hand, it was established in [9] that, for this same class of systems, the origin is globally asymptotically stabilizable by nonlinear feedback. A particular algorithm using multiple saturation functions was initiated in [13] and completed in [11].

On a slightly different front, there has been renewed interest in the domains of attraction that can be achieved using linear feedback. In a pair of recent papers, Lin and Saberi have shown that, both for discrete time [4] and continuous time [5], the domain of attraction for a linear system null controllable with bounded controls and subject to input saturation can be made arbitrarily large using appropriately tuned linear feedback. The purpose of this note is to provide a simple Lyapunov proof for the result of [5] which, at the same time, allows for some extensions. In addition, it is illustrated with an example that the result of this note is useful for semi-global stabilization of some nonlinear systems that appear unrelated to the problem of stabilization with input saturation.

The particular choice of feedback that is used to achieve the extensions of this note is based on the solution to a family of "full information" H_∞ control problems [2]. The solution to this problem provides a family of arbitrarily small feedback gains and a corresponding family of Lyapunov functions. Given a prescribed compact set, a member of this feedback family is selected which exhibits stability robust to input nonlinearities for initial conditions in this compact set.

II. PROBLEM STATEMENT

We initially consider systems of the form:

$$\begin{aligned} \dot{x} &= Ax + \sigma(u) \\ y &= C'x, \end{aligned} \quad (1)$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$ and $\sigma: \mathbb{R}^m \rightarrow \mathbb{R}^m \subset \mathbb{R}^m$ where Γ is some bounded neighborhood of the origin and $\sigma(0) = 0$. The objective is to find a family of linear controllers that can be used to make the domain of attraction for the point $x = 0$ arbitrarily large. This problem is referred to as the semi-global stabilization problem:

Definition 2.1: The point $x = 0$ of (1) is semi-globally stabilizable by linear state feedback if, for each compact set U , there exists a feedback $u = Kx$ such that the origin is a locally asymptotically stable equilibrium with domain of attraction containing U .

Definition 2.2: The point $x = 0$ of (1) is semi-globally stabilizable by linear dynamic output feedback if, for each compact set U , there exist a dynamic output feedback of the form

$$\begin{aligned} \dot{\zeta} &= F\zeta + G y \\ u &= K\zeta \end{aligned} \quad (2)$$

and a compact set U_ζ such that the point $(x, \zeta) = (0, 0)$ is a locally asymptotically stable equilibrium with domain of attraction containing $U \times U_\zeta$.

We make the following assumptions:

Assumption 1: The linear approximation of (1) exists and is stabilizable with bounded controls, i.e., the pair $(A, \sigma'(0))$ is stabilizable, and the eigenvalues of A have nonpositive real part.

Assumption 2: The pair (C', A) is detectable.

The next section is dedicated to a Lyapunov proof for the following result:

Theorem 1: If Assumption 1 is satisfied then the origin of (1) is semi-globally stabilizable by linear state feedback. If Assumptions 1 and 2 are satisfied then the origin of (1) is semi-globally stabilizable by linear dynamic output feedback.

Remark 2.1 Lin and Saberi have proved this result in [5] under the additional assumptions that $\sigma(u) = B\bar{\sigma}(u)$ where (A, B) is stabilizable, the i th entry of the vector $\bar{\sigma}(u)$ depends only on u_i and each entry is linear near the origin, and the nonlinear element σ is used in the compensator (2).

To motivate the solution to this problem, we note that, for each $\rho > 0$, $\exists \Delta > 0$ such that

$$|\sigma(u) - \sigma'(0)u| \leq \rho|u| \quad \forall u \in \mathbb{R}^m, |u| \leq \Delta. \quad (3)$$

This bound can be exploited if $|u|$ can be kept sufficiently small. For a linear system of the form

$$\dot{x} = Ax + \sigma'(0)u \quad (4)$$

satisfying Assumption 1, for any compact set of initial conditions, a linear stabilizing feedback can be chosen which makes $\sup_{t \geq 0} |u(t)|$ arbitrarily small. This was shown in [5] for example. The contribution of this note is that feedbacks with this same property can be constructed from the solution to an appropriate full information H_∞ control problem, thereby achieving stability robust to perturbations of the form $\sigma(u) - \sigma'(0)u$ satisfying the bound (3) for ρ sufficiently small.

III. PROOF OF THEOREM 1

We will solve a more general problem than the one considered in Section II. Consider the system:

$$\dot{x} = Ax + B_1 u + B_2 g(u). \quad (5)$$

We recover (1) by taking $B_1 = \sigma'(0)$, $B_2 = I$ and $g(u) = \sigma(u) - \sigma'(0)u$. Then Assumption 1 becomes:

Assumption 1b: The pairs (A, B_1) and (A, B_2) are stabilizable and the eigenvalues of A have nonpositive real part.

We will form a family of semi-globally stabilizing feedbacks for (5) from the following result:

Lemma 3.1. Let Assumption 1b hold. If $Q: (0, 1] \rightarrow \mathbb{R}^{n \times n}$ is a continuous, positive definite matrix valued function and

$$\lim_{\epsilon \rightarrow 0} \lambda_{\max}(Q(\epsilon)) = 0 \quad (6)$$

then there exists a positive real number γ such that, for each $\epsilon \in (0, 1]$, $\exists P(\epsilon) > 0$ satisfying

$$\begin{aligned} 1) \quad & A'P(\epsilon) + P(\epsilon)A + P(\epsilon)\left(\frac{1}{\gamma^2}B_2B_2' - B_1B_1'\right)P(\epsilon) + Q(\epsilon) = 0 \\ 2) \quad & \end{aligned} \quad (7)$$

$$\lim_{\epsilon \rightarrow 0} \lambda_{\max}(P(\epsilon)) = 0. \quad (8)$$

Proof: Using the results of [2], since Q is positive definite for each $\epsilon \in (0, 1]$ and (A, B_1) and (A, B_2) are stabilizable pairs, for each $\epsilon \in (0, 1]$ there exists $\gamma_{opt}(\epsilon)$ such that, for each $\gamma > \gamma_{opt}(\epsilon)$, there exists $P(\epsilon, \gamma) > 0$ satisfying

$$A^T P + P A + P \left(\frac{1}{\gamma^2} B_2 B_2^T - B_1 B_1^T \right) P + Q(\epsilon) = 0. \quad (9)$$

In addition, since

$$\sup_{\epsilon \in (0, 1]} \lambda_{\max}(Q(\epsilon)) < \infty \quad (10)$$

we can choose γ to satisfy

$$\gamma > \bar{\gamma} := \sup_{\epsilon \in (0, 1]} \gamma_{opt}(\epsilon) \quad (11)$$

so that, for each $\epsilon \in (0, 1]$, there exists $P(\epsilon) > 0$ satisfying (7).

To establish (8), we first note that, for each $\epsilon \in (0, 1]$, there exists $X_2(\epsilon) > 0$ satisfying

$$A^T X_2(\epsilon) + X_2(\epsilon) A - X_2(\epsilon) B_1 B_1^T X_2(\epsilon) + Q(\epsilon) = 0 \quad (12)$$

and

$$\lim_{\epsilon \rightarrow 0} \lambda_{\max}(X_2(\epsilon)) = 0. \quad (13)$$

This follows from a result of [16] which gives that, since the eigenvalues of A have nonpositive real part, $X_2 = 0$ is the unique positive semi-definite solution to

$$A^T X_2 + X_2 A - X_2 A - X_2 B_1 B_1^T X_2 = 0 \quad (14)$$

and the result of [15] on continuity of the solution to the algebraic Riccati equation (12) under the condition (6). Second, using (11) and [6, Theorem 3.1], we have that, for each $\epsilon \in (0, 1]$,

$$P(\epsilon) \leq \frac{\gamma}{\gamma^2 - \bar{\gamma}^2} X_2(\epsilon) \quad (15)$$

where $P(\epsilon)$ and $X_2(\epsilon)$ are the solutions to the Riccati equations (7) and (12), respectively. The bound (15) can also be established with the tools of [2, Section VI.A, VII.C]. Then (8) follows from (13) and (15).

A. The State Feedback Result

Assumption 3: There exist positive definite matrix valued functions $Q: (0, 1] \rightarrow \mathbb{R}^{n \times n}$ and $P: (0, 1] \rightarrow \mathbb{R}^{n \times n}$ and strictly positive real numbers γ and Δ satisfying (6)–(8) and

$$|g(u)| \leq \frac{1}{\Delta}, \quad \forall u \in \{u \in \mathbb{R}^m : |u| \leq \Delta\}. \quad (16)$$

Remark 3.1: There is no requirement that g be differentiable at the origin. All that is required is that $g(0) = 0$ and g is locally Lipschitz at the origin with a sufficiently small Lipschitz constant.

From (3) and Lemma 3.1, Assumption 3 holds if Assumption 1 holds. Thus, the state feedback result of Theorem 1 is a consequence of the following result:

Proposition 3.1: If Assumption 3 holds then the family of linear state feedbacks:

$$u = -B_1^T P(\epsilon)x, \quad (17)$$

parameterized by ϵ , semi-globally stabilizes the origin of (5).

Proof: Define $\nu = \lambda_{\max}(\sqrt{B_1 B_1^T})$. Let U be a compact set such that $x(0) \in U$. Let c be a strictly positive real number such that

$$c^2 \geq \sup_{\substack{x \in U \\ \epsilon \in (0, 1]}} \{x^T P(\epsilon)x\} \quad (18)$$

and pick $\epsilon^* \in (0, 1]$ so that

$$\lambda_{\max}(P(\epsilon^*)) \leq \left(\frac{\Delta}{c\nu} \right)^2. \quad (19)$$

Since (8) holds, c and ϵ^* exist. Define $z = P^{\frac{1}{2}}(\epsilon^*)x$ so that $u = -B_1^T P^{\frac{1}{2}}(\epsilon^*)z$. From (19) it follows that

$$(20)$$

In the z coordinates, the closed loop is given by

$$\begin{aligned} \dot{z} &= P^{\frac{1}{2}}[AP^{-\frac{1}{2}}z - B_1 B_1^T P^{\frac{1}{2}}z + B_2 g(u(z))] \\ &= A_z z + P^{\frac{1}{2}} B_2 g(u(z)). \end{aligned} \quad (21)$$

Note from (8) that

$$A_z^T + A_z = -\frac{1}{\gamma^2} P^{\frac{1}{2}} B_2 B_2^T P^{\frac{1}{2}} - P^{\frac{1}{2}} B_1 B_1^T P^{\frac{1}{2}} - P^{-\frac{1}{2}} Q P^{-\frac{1}{2}}. \quad (22)$$

Let $V(z) = z^T z$ and consider \dot{V} on the set

$$D = \{z : V(z) \leq c^2\}. \quad (23)$$

From (16) and (20) it follows that

$$|g(u(z))| \leq \frac{1}{\Delta} |u(z)| \quad \forall z \in D. \quad (24)$$

Therefore, $\forall z \in D$,

$$\begin{aligned} \dot{V} &= -z^T \left(\frac{1}{\gamma^2} P^{\frac{1}{2}} B_2 B_2^T P^{\frac{1}{2}} + P^{\frac{1}{2}} B_1 B_1^T P^{\frac{1}{2}} + P^{-\frac{1}{2}} Q P^{-\frac{1}{2}} \right) z \\ &\quad + 2z^T P^{\frac{1}{2}} B_2 g(u(z)) \\ &\leq -\frac{1}{\gamma^2} |B_2^T P^{\frac{1}{2}} z|^2 - |B_1^T P^{\frac{1}{2}} z|^2 - |Q^{\frac{1}{2}} P^{-\frac{1}{2}} z|^2 \\ &\quad + \frac{2}{\gamma} |B_2^T P^{\frac{1}{2}} z| |B_1^T P^{\frac{1}{2}} z| \\ &\leq -|Q^{\frac{1}{2}} P^{-\frac{1}{2}} z|^2. \end{aligned} \quad (25)$$

Since Q is positive definite, (25) and the definition of V imply that every initial condition in the set D converges exponentially to the origin. From (18) and (23), $x \in U \Rightarrow z \in D$.

B. The Output Feedback Result

Assumption 3b: There exist positive definite matrix valued functions $Q: (0, 1] \rightarrow \mathbb{R}^{n \times n}$ and $P: (0, 1] \rightarrow \mathbb{R}^{n \times n}$, strictly positive real numbers γ and Δ and a matrix of gains L such that

- 1) (6)–(8) are satisfied,
- 2) $A + LC$ is Hurwitz
- 3) defining:

- a) $\beta = \lambda_{\max}(P_\alpha)$ where $(A + LC)^T P_\alpha + P_\alpha (A + LC) = -I$,
- b) $\nu = \sqrt{\lambda_{\max}(B_1 B_1^T)}$,
- c) $\kappa = \sup_{\epsilon \in (0, 1]} \lambda_{\max}(P(\epsilon))$,

- d) $\mu = \sqrt{6} \max\{\beta(2 + 4v^2\kappa^2), \gamma\}$,
 e) we have

$$|g(u)| \leq \frac{1}{\mu}|u| \quad \forall u \in \{u \in \mathbb{R}^m : |u| \leq \Delta\}. \quad (26)$$

From (3) and Lemma 3.1, Assumption 3b holds if Assumptions 1 and 2 hold. Thus, the output feedback result of Theorem 1 follows from the result:

Proposition 3.2: If Assumption 3b holds then the family of dynamic, linear output feedbacks:

$$\begin{aligned} \dot{\hat{x}} &= A\hat{x} + B_1 u + L(C\hat{x} - y) \\ u &= -B_1^T P(\epsilon)\hat{x} \end{aligned} \quad (27)$$

parameterized by ϵ , \hat{x} initialized in a compact subset of \mathbb{R}^n , semi-globally stabilizes the origin of (5).

Proof: With P_o , v and κ defined in Assumption 3b, define

$$\begin{aligned} \lambda_{\min}(P_o) \\ \tau &= 2 + 4v^2\kappa^2. \end{aligned} \quad (28)$$

Let U and U_x be compact sets such that $(x(0), \hat{x}(0)) \in U \times U_x$ and let c be a strictly positive real number such that

$$c^2 > \sup \{x^T P(\epsilon)x + \tau(x - \hat{x})^T P_o(x - \hat{x})\}. \quad (29)$$

Pick $\epsilon^* \in (0, 1]$ so that

$$\lambda_{\max}(P(\epsilon^*)) \leq \min \left\{ \left(\frac{\Delta}{2c\tau} \right)^2, \sqrt{\alpha\tau}\Delta \right\} \quad (30)$$

Define $e = x - \hat{x}$ and $z = P^{\frac{1}{2}}(\epsilon^*)e$. Observe that

$$u = -B_1^T P^{\frac{1}{2}}z + B_1^T P e \quad (31)$$

and, from the definition of v in Assumption 3b and (30), it follows that

$$|u| \leq \frac{\Delta}{2c}(|z| + \sqrt{\alpha\tau}|e|). \quad (32)$$

In the coordinates e , z the dynamics are given by

$$\begin{aligned} \dot{e} &= A_e e + B_2 g(u) \\ \dot{z} &= A_z z + P^{\frac{1}{2}} B_1 B_1^T P e + P^{\frac{1}{2}} B_2 g(u), \end{aligned} \quad (33)$$

where A_z is defined in (21). Choose the Lyapunov function candidate

$$V(z, e) = z^T z + \tau e^T P_o e \quad (34)$$

and consider \dot{V} on the set

$$\mathcal{D} = \{(z, e) : V(z, e) \leq c^2\} \quad (35)$$

with c defined in (29). Observe that

$$\begin{aligned} \max_{(z,e) \in \mathcal{D}} |z| &= c \\ \max_{(z,e) \in \mathcal{D}} |e| &= c(\sqrt{\tau\alpha})^{-1}. \end{aligned} \quad (36)$$

Then, from (26) and (32),

$$|g(u)| \leq \frac{1}{\mu} \quad \forall (z, e) \in \mathcal{D}. \quad (37)$$

Also, from the definition of v and κ in Assumption 3b,

$$|B_1^T P e| \leq v\kappa|e|. \quad (38)$$

Therefore, $\forall (z, e) \in \mathcal{D}$,

$$\begin{aligned} \dot{V} &= -\tau|e|^2 + 2\tau e^T P_o B_2 g(u) \\ &\quad - z^T \left(\frac{1}{\gamma^2} P^{\frac{1}{2}} B_2 B_2^T P^{\frac{1}{2}} + P^{\frac{1}{2}} B_1 B_1^T P^{\frac{1}{2}} + P^{-\frac{1}{2}} Q P^{-\frac{1}{2}} \right) z \\ &\quad + 2z^T P^{\frac{1}{2}} B_1 B_1^T P e + 2z^T P^{\frac{1}{2}} B_2 g(u) \\ &\leq -\tau|e|^2 + \frac{2\tau\beta}{\mu}|e|(|B_1^T P^{\frac{1}{2}}z| + v\kappa|e|) \\ &\quad - \frac{1}{\gamma^2}|B_2^T P^{\frac{1}{2}}z|^2 - |B_1^T P^{\frac{1}{2}}z|^2 - |Q^{\frac{1}{2}} P^{-\frac{1}{2}}z|^2 \\ &\quad + 2|B_1^T P^{\frac{1}{2}}z|(|v\kappa|e|) + \frac{2}{\mu}|B_2^T P^{\frac{1}{2}}z|(|B_1^T P^{\frac{1}{2}}z| + v\kappa|e|). \end{aligned} \quad (39)$$

Using, for all real numbers a and b ,

$$ab \leq \frac{1}{3}a^2 + \frac{2}{3}b^2 \quad (40)$$

we get

$$\begin{aligned} &\frac{2\tau\beta}{\mu}|e|(|B_1^T P^{\frac{1}{2}}z| + v\kappa|e|) \\ &\leq \frac{6\tau^2\beta^2}{\mu^2}|e|^2 + \frac{1}{3}|B_1^T P^{\frac{1}{2}}z|^2 + \frac{1}{3}v^2\kappa^2|e|^2 \\ &\quad 2|B_1^T P^{\frac{1}{2}}z|(|v\kappa|e|) \\ &\leq \frac{1}{3}|B_1^T P^{\frac{1}{2}}z|^2 + 3v^2\kappa^2|e|^2 \\ &\quad \frac{2}{\mu}|B_2^T P^{\frac{1}{2}}z|(|B_1^T P^{\frac{1}{2}}z| + v\kappa|e|) \\ &\leq \frac{6}{\mu^2}|B_2^T P^{\frac{1}{2}}z|^2 + \frac{1}{3}|B_1^T P^{\frac{1}{2}}z|^2 + \frac{1}{3}v^2\kappa^2|e|^2. \end{aligned} \quad (41)$$

It follows that, $\forall (z, e) \in \mathcal{D}$,

$$\begin{aligned} \dot{V} &\leq -\tau|e|^2 + \frac{6\tau^2\beta^2}{\mu^2}|e|^2 + 4v^2\kappa^2|e|^2 - \frac{1}{\gamma^2}|B_2^T P^{\frac{1}{2}}z|^2 \\ &\quad + \frac{6}{\mu^2}|B_2^T P^{\frac{1}{2}}z|^2 - |Q^{\frac{1}{2}} P^{-\frac{1}{2}}z|^2. \end{aligned} \quad (42)$$

Then, from (28) and the definition of μ in Assumption 3b,

$$\dot{V} \leq -|e|^2 - |Q^{\frac{1}{2}} P^{-\frac{1}{2}}z|^2 \quad \forall (z, e) \in \mathcal{D}. \quad (43)$$

Since Q is positive definite, (43) and (34) imply that every initial condition in the set \mathcal{D} converges exponentially to the origin. Finally, from (29) and (35), $(x, \hat{x}) \in U \times U_x \Rightarrow (z, e) \in \mathcal{D}$.

IV. APPLICATION: A SYSTEM NONLINEAR IN THE STATE

The result of this note can be used as a tool for semi-globally stabilizing the origin of the system

$$\begin{aligned} \dot{x}_1 &= x_2 + x_3^2 \\ \dot{x}_2 &= x_3 \\ \dot{x}_3 &= u \end{aligned} \quad (44)$$

for example, using linear feedback. A nonlinear globally stabilizing solution is given in [14]. Indeed, semi-global stabilization by linear feedback can be accomplished in two steps. First, using the results described here, the origin of

$$\begin{aligned} \dot{x}_1 &= x_2 + v^2 \\ \dot{x}_2 &= v \end{aligned} \quad (45)$$

can be semi-globally stabilized using linear state feedback, $v = -p_{21}(\epsilon)x_1 - p_{22}(\epsilon)x_2$. Moreover, a Lyapunov function that demonstrates the result is available. Then, using the semi-global backstepping tool of [12], the origin of (44) can be semi-globally stabilized

with a linear state feedback of the form

$$u = -K(x_3 - p_{21}(\epsilon)x_1 + p_{22}(\epsilon)x_2). \quad (46)$$

Likewise, with the output feedback result of this note, the origin of (44) is semi-globally stabilizable by linear dynamic feedback with x_1 and x_3 as measurements.

V. CONCLUSION

A solution to the problem of semi-global stabilization by linear feedback for linear systems with input saturation has been proposed which is an alternative to the solution given in [5]. The approach used is based on the solution to a family of H_∞ Riccati equations. As compared with the result in [5], the semi-global result is achieved while significantly relaxing the requirements on the input characteristic. Moreover, an output feedback solution is derived which is completely linear and hence does not explicitly use the input characteristic. On the other hand, we have not attempted to make any comparison with [5] regarding numerical and implementation issues.

It has also been demonstrated that this result is useful for semi-global stabilization of some nonlinear systems that would otherwise seem unrelated to the problem of stabilization with input saturation.

ACKNOWLEDGMENTS

The author would like to thank Z. Lin, A. Saberi and A. Packard for pointing out references which were used to shorten the original proof of Lemma 3.1.

REFERENCES

- [1] B. D. O. Anderson and J. B. Moore, *Linear Optimal Control*, Englewood Cliffs: Prentice-Hall, 1971.
- [2] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, "State-space solutions to standard H_2 and H_∞ control problems," *IEEE Trans. Automat. Contr.*, vol. 34, no. 8, pp. 831-847, 1989.
- [3] A. T. Fuller, "In the large stability of relay and saturated control systems with linear controllers," *Int. J. Contr.*, vol. 10, pp. 457-480, 1969.
- [4] Z. Lin and A. Saberi, "Semi-global exponential stabilization of linear discrete-time systems subject to 'input saturation' via linear feedbacks," Submitted to *Syst. Contr. Lett.*, 1993.
- [5] —, "Semi-global exponential stabilization of linear systems subject to 'input saturation' via linear feedbacks," *Syst. Contr. Lett.*, vol. 21, no. 3, pp. 225-239, 1993.
- [6] M. A. Rotea and A. E. Frazho, "Bounds on solutions to the H_∞ algebraic Riccati equations and the H_2 properties of H_∞ central solutions," *Syst. Contr. Lett.*, vol. 19, no. 5, pp. 341-352, 1992.
- [7] W. E. Schmitendorf and B. R. Barmish, "Null controllability of linear systems with constrained controls," *SIAM J. Control Optim.*, vol. 18, pp. 327-345, 1980.
- [8] E. D. Sontag, "An algebraic approach to bounded controllability of linear systems," *Int. J. Contr.*, vol. 39, pp. 181-188, 1984.
- [9] E. D. Sontag and H. J. Sussmann, "Nonlinear output feedback design for linear systems with saturating controls," in *Proc. 29th IEEE Conf. Decis. Contr.*, Dec. 1990, pp. 3414-3416.
- [10] H. J. Sussmann and Y. Yang, "On the stabilizability of multiple integrators by means of bounded feedback controls," in *Proc. 30th IEEE Conf. Decis. Contr.*, Dec. 1991, pp. 70-73.
- [11] H. J. Sussmann, E. D. Sontag, and Y. Yang, "A general result on the stabilization of linear systems using bounded controls," submitted to *IEEE Trans. Automat. Contr.*, 1993.
- [12] A. R. Teel and L. Praly, "Tools for semi-global stabilization by partial state and output feedback," *SIAM J. Contr. Optim.*, to appear.
- [13] A. R. Teel, "Global stabilization and restricted tracking for multiple integrators with bounded controls," *Syst. Contr. Lett.*, vol. 18, no. 3, pp. 165-171, 1992.
- [14] —, "Using saturation to stabilize single-input partially linear composite nonlinear systems," in *Proc. IFAC Nonlinear Control Systems Design Symposium*, June 1992, pp. 224-229.
- [15] H. L. Trentelman, "Families of linear-quadratic problems: Continuity properties," *IEEE Trans. Automat. Contr.*, vol. 32, no. 4, pp. 323-329, 1987.
- [16] J. C. Willems, "Least squares stationary optimal control and the algebraic Riccati equation," *IEEE Trans. Automat. Contr.*, vol. AC-16, no. 6, pp. 621-634, 1971.

Stability and Exponential Stability of an Adaptive Control Scheme for Plants of any Relative Degree

Giorgio Bartolini, Antonella Ferrara, and Alexander A. Stotsky

Abstract—Relying on a new adaptive pole assignment control scheme for plants with unknown relative degree and uncertain order which can be made minimum phase by means of a relative-degree-one first order parallel compensator, a new adaptation mechanism is proposed in this note in order to tackle with the problem of the attainment of an exponentially stable behavior of the output error. In particular, the connection between exponential stability and persistent excitation is outlined, and a complete stability analysis is performed.

I. INTRODUCTION

Starting from the works by Landau, Monopoli, and Narendra [1], [2], [3], the research activity relevant to the model reference approach to the adaptive control of LTI SISO systems has been mainly oriented towards both the simplification of the basic schemes and the modification of the adaptation mechanisms to cope with the nonstandard cases of unmodeled dynamics and bounded disturbances [4]-[7]. The common target has been however represented by the necessity of assuring, at least, the boundedness of tracking and parameter errors.

Recently, the attention has been focused on the development of adaptive schemes for plants with unknown relative degree. In [7], for instance, the relative degree is uncertain but its uncertainty is tied to the uncertainty on the plant description which is expressed in multiplicative as well as additive form however required to be small enough. While in [8] and [9], a simplified adaptive control scheme has been presented, which performs the regulation of uncertain plants via pole assignment without requiring the perfect knowledge of the relative degree of the plant to be controlled and independently of the magnitude of the unmodeled dynamics. Yet, in this simplified scheme, traditional integral adaptation mechanisms have been adopted. Consequently, no great emphasis has been placed on the analysis of transient behaviors, as usually happens when only asymptotic stability is ensured.

Manuscript received October 19, 1992; revised March 23, 1994.

G. Bartolini and A. Ferrara are with the Department of Communication, Computer and System Sciences, University of Genova, Via Opera Pia 11A, 16145 Genova, Italy.

A. A. Stotsky is with the Institute for Problems of Mechanical Engineering, Academy of Sciences of Russia, Lensoveta st. 57-32, St. Petersburg 196143, Russia.

IEEE Log Number 9407017.

As a matter of fact, the convergence rate of the adjustable parameters when using traditional integral adaptation mechanism can be slow even under the condition of persistency of excitation and can be improved by the introduction of a prediction error into the adaptation law. Controllers composed of the terms driven by both the tracking and the prediction error for robot manipulators were proposed by Slotine and Li [10]. To avoid the measurement of the derivatives of the tracking error when getting the prediction error, various types of filtering have been proposed [6], [11], [12]. However, in previous works, the term which is responsible for filter design was not included in the Lyapunov function.

The design procedure for constructing a prediction error used in this note is based on the Lyapunov function method and allows one to perform the filter design simultaneously with algorithm design, to connect the parameters of filter with algorithm parameters and to include a relay term driven by the estimate of the prediction error in the filter for robustness enhancement, in accordance with [13].

The main contribution of this note is the introduction of a new combined direct and indirect adaptive algorithm, driven by both the tracking and the prediction error, which does not require the perfect knowledge of the relative degree of the plant. Moreover, exponential stability of both the tracking and the parameter error is established when some persistent excitation requirements are met.

The present work is organized as follows. In the next section the existence of a control structure characterized by a complexity which is independent of both the relative degree and the precise knowledge of the order is first discussed. Section III is devoted to the construction of a suitable error model relevant to the adaptive case. Finally, the description of the new adaptation mechanism and the analysis of the stability properties of the overall adaptive scheme is provided in Section IV.

II. THE UNDERLYING LINEAR STRUCTURE OF THE CONTROLLER

We consider the following LTI SISO plant

$$\begin{aligned} \dot{x}_p(t) &= \Lambda_p x_p(t) + b_p u_p(t) \\ y_p(t) &= h_p^T x_p(t) \end{aligned} \quad (1)$$

where $x_p(t) \in \mathcal{R}^q$, with $n_1 \leq q \leq n_2$, $u_p(t) \in \mathcal{R}^1$, Λ_p is a matrix of suitable dimension, and, using the operational notation $s = d/dt$,

$$y_p(t) = h_p^T (sI - \Lambda_p)^{-1} b_p = k_p \frac{B(s)}{A(s)} u_p(t). \quad (2)$$

This plant is assumed to satisfy the following assumptions: a) $A(s)$, $B(s)$ are coprime not necessarily Hurwitz polynomials; b) $\deg(B(s)) \leq q - 1$; c) the parameters of Λ_p , b_p , h_p are supposed to be unknown; d) only the input $u_p(t)$ and the output $y_p(t)$ are assumed available; e) the sign of the gain k_p is known. Note that this set of assumptions allows us to deal with the case in which the reference model order is lower than the plant order. In particular, to simplify the treatment, we shall make use of a first order strictly proper model, although the plant relative degree and order are not exactly known as happens in any general representation of systems with unmodeled dynamics [6].

The control problem we are dealing with can be stated as follows.

Problem 1: Find a control law $u_p = u_p(t, y_p(t), r(t))$ such that for any possible value of the plant order q

$$y_p(t) = k_p \frac{B(s)}{A_q(s)} r(t), \quad (3)$$

where $r(t)$ is a suitable reference input which is assumed to be bounded, and $A_q(s)$ is a polynomial with degree equal to q belonging to a set of arbitrarily chosen Hurwitz polynomials with degree between n_1 and n_2 (in practice, a pole assignment control problem has to be solved in spite of the uncertainty we assume on the plant).

To face the control problem above introduced, the following parameterization of the control law $u_p(t)$ can be chosen

$$u_p(t) = \Theta^{*T} X(t), \quad (4)$$

where $X^T(t) = [x_1^T(t) y_a(t) x_2^T(t) r(t) x_3^T(t)]$ is the vector containing the signals of interest. Among them, the vector signals $x_i(t)$ are the states of three state variable filters described by the following differential equations

$$\begin{aligned} \dot{x}_1(t) &= \Lambda x_1(t) + b_1 u_p(t) \\ \dot{x}_2(t) &= \Lambda x_2(t) + b_1 y_a(t) \\ \dot{x}_3(t) &= \Lambda x_3(t) + b_1 r(t), \end{aligned} \quad (5)$$

where Λ is a $n_2 \times n_2$ stable matrix, and $y_a(t)$ is an auxiliary signal obtained as $y_a(t) = y_p(t) + y_r(t)$, $y_r(t)$ being the output signal generated by a rational feedforward compensator, namely,

$$\dot{y}_r(t) = -a y_r(t) + k_1 u_p(t), \quad (6)$$

(k_1 and a positive constants). Hence, $y_a(t)$ can be regarded as the output of the parallel between the plant and a first order relative-degree-one filter. Finally, Θ^* is a constant vector of suitable dimension, containing the parameters which solve the control problem. The introduction of the parallel compensator (6) is motivated by the following considerations. The relative degree of the parallel connection of the plant and the first order filter turns out to be one regardless of the relative degree and of the order of the true plant. Indeed,

$$\begin{aligned} y_a(t) &= k_p \frac{B(s)}{A(s)} u_p(t) + \frac{k_1}{s+a} u_p(t) \\ &= \frac{k_1 A(s) + k_p B(s)(s+a)}{A(s)(s+a)} u_p(t). \end{aligned} \quad (7)$$

Note that in this work we assume that a known parallel compensator $H(s) = N_r(s)/D_r(s)$, with relative degree equal to one, such that $N_r(s)A(s) + k_p D_r(s)B(s)$ is Hurwitz exists. This means that the class of systems to which the proposed approach is applicable is enlarged with respect to classical MRAC schemes to contain those nonminimum phase systems whose zeros can be robustly located in the complex l.h.p. by the parallel connection with a known relative-degree-one compensator [14]. In particular, in [14] the conditions under which the problem is solvable by means of a first order parallel compensator are reported. While, with reference to systems for which the above robust stabilization problem cannot be solvable, a possible solution to the control problem has been proposed in [15]. However, this topic is still under investigation (see [16] for some preliminary results).

More precisely, in writing (6)–(7), for the sake of simplicity we are implicitly assuming that $H(s) = 1/(s+a)$, (i.e., $k_1 = 1$) suffices to satisfy the minimum phase requirement for the roots of the numerator of the transfer function in (7). On the basis of these considerations, it is possible to claim that, by means of the parallel compensator, an auxiliary minimum phase and relative-degree-one plant has been obtained. Then, we are in a position to cope with the relevant control problem previously stated. To this end, let us introduce the following results.

Theorem 1: For any minimum phase and relative-degree-one plant represented by the input output pair $\{u_p(t), y_p(t)\}$ whose order can range between two known values $[n_1, n_2]$, there exists a parameter vector pair θ_1^*, θ_2^* of fixed dimensions, and a scalar value θ_{20}^* , such that, by using a control law of the type

$$u_p^*(t) = -\theta_1^{*T} x_1(t) - \theta_2^{*T} x_2(t) - \theta_{20}^* y_p(t) + w^*(t), \quad (8)$$

where $x_1(t), x_2(t) \in \mathcal{R}^{n_2}$, and $w^*(t)$ is a signal depending on the reference input, as will be specified below, a unique transfer function $T_m(s)$ exists such that the following condition is satisfied

$$y_p(t) = y_n^*(t) = T_m(s)w^*(t) \quad (9)$$

$T_m(s)$ being a strictly positive real (SPR) transfer function.

Proof: If $D(s)$ is the arbitrarily chosen characteristic polynomial of degree equal to n_2 associated with matrix Λ in (5), and $F_1(s), F_2(s)$ are the numerators of $\theta_1^{*T}(sI - \Lambda)^{-1}b_j$ and $\theta_2^{*T}(sI - \Lambda)^{-1}b_j + \theta_{20}^*$, respectively, denoting with $\tilde{B}(s)/\tilde{A}(s)$ the transfer function of the parallel structure (auxiliary plant) indicated in (7), then the transfer function between $w^*(t)$ and $y_p(t)$ results

$$T_m(s) = \frac{\tilde{B}(s)D(s)}{\tilde{A}(s)(D(s) - F_1(s)) - F_2(s)\tilde{B}(s)} \quad (10)$$

The model transfer function $T_m(s)$ needs to be independent of the considered auxiliary plant $\tilde{B}(s)/\tilde{A}(s)$. Keeping in mind this fact, the proof can be developed by introducing the following factorization $D(s) = \prod_{i=1}^q (s + d_i) = \prod_{i=q+1}^{n_2} (s + d_i) \prod_{j=1}^q (s + d_j) = D'(s)D''(s)$ and choosing

$$\begin{aligned} F_1(s) - D(s) &= D'(s)\tilde{B}(s) \\ F_2(s) &= D'(s)(\tilde{A}(s) - \hat{A}_q(s)), \end{aligned} \quad (11)$$

where the monic polynomial $\hat{A}_q(s)$, with the same degree as $\tilde{A}(s)$, can be expressed as $\hat{A}_q(s) = (s + a)A_q(s)$, and $A_q(s)$ belongs to a fixed set of polynomials of the type $\{A_k(s); A_{k+1}(s) = (s + \lambda_k)A_k(s), k = n_1, \dots, n_2\}$, with $A_{n_1}(s)$ and $\{\lambda_k\}$ chosen so that for any q , $A_q(s)$ is the relevant characteristic polynomial to be assigned to the plant.

From (10) and (11), $T_m(s)$ results equal to $D''(s)/\hat{A}_q(s)$, and can be invariant for any q only if $D''(s)$ and $\hat{A}_q(s)$ have common factors. Due to the arbitrariness of the choice of $D(s)$ and $A_k(s)$, we can always assume that this is the case. Therefore, $T_m(s)$ can be rewritten as

$$T_m(s) = \frac{D_m(s)}{A_m(s)} \quad (12)$$

where $A_m(s) := \hat{A}_{n_1}(s)$. Note that in (12) the following factorization has been used

$$\hat{A}_q(s) = A_m(s)D'''(s) \quad D''(s) = D_m(s)D'''(s) \quad (13)$$

Hence, the same $T_m(s)$ is obtained starting from any admissible plant, and polynomials $F_1(s), F_2(s)$, and $D(s)$ are of constant degree (so that θ_1^* and $[\theta_{20}^* \theta_2^{*T}]^T$, which contain the coefficients of $F_1(s)$ and $F_2(s)$, are of fixed dimensions) even if the degree of $\tilde{A}(s)$ ranges in the interval $[n_1 + 1, n_2 + 1]$, which completes the proof. \square

Theorem 2: In the same conditions as in Theorem 1, there exists a parameter vector $[\theta_{30}^* \theta_3^{*T}]^T$, such that the choice

$$w^*(t) = \theta_{30}^{*T} x_3(t) + \theta_3^{*T} r(t) \quad (14)$$

ensures the complete satisfaction of the control object (3), for any plant belonging to the set of systems fulfilling assumptions a)–d).

Proof: The transfer function between $w^*(t)$ and $y_p(t)$, due to the choice of the control law indicated in (8), results

$$y_p(t) = T_m(s) \frac{k_p B(s)(s + a)}{\tilde{B}(s)} w^*(t), \quad (15)$$

where, for the sake of simplicity, $T_m(s) = D_m(s)/A_m(s)$ can be chosen as $T_m(s) = 1/(s + a)$, i.e., $A_m(s) = (s + a)D_m(s)$. By posing $w^*(t) = (F_3(s)/D(s))r(t)$, where the coefficients of polynomial $F_3(s)$ coincide with the components of the vector $[\theta_{30}^* \theta_3^{*T}]^T$, we obtain

$$y_p(t) = \frac{k_p B(s)F_3(s)}{\tilde{B}(s)D(s)} r(t). \quad (16)$$

Hence, choosing

$$F_3(s) = \tilde{B}(s)D'(s), \quad (17)$$

where $D'(s)$ has been specified above, it results

$$y_p(t) = \frac{k_p B(s)}{D''(s)} r(t). \quad (18)$$

Taking into account (10)–(11), the definition of $A_q(s)$ and the choice above made for $T_m(s)$, one has $T_m(s) = D''(s)/[(s + a)A_q(s)] = 1/(s + a)$. Therefore, $D''(s) = A_q(s)$ which is the relevant characteristic polynomial to be assigned. So the pole assignment requirement is fulfilled and Theorem 2 is proved. \square

It should be explicitly noted that the number of parameter sets to be tuned is only two (namely, the sets of coefficients of polynomials $F_1(s)$ and $F_2(s)$), because of the relationship between $F_1(s)$ and $F_3(s)$ expressed by (11) and (17).

Indeed,

$$\frac{F_1(s)}{D(s)} = \theta_{30}^* + \frac{F_3(s) - D(s)}{D(s)} \quad (19)$$

($\theta_{30}^* = 1$, since $F_3(s)$ and $D(s)$ are monic), and $F_3(s) - D(s) = F_1(s)$, so that

$$\theta_1^* = \theta_1^* + \underline{d} \quad (20)$$

\underline{d} being the vector of the known coefficients of $D(s)$. Moreover, with respect to classical adaptive control schemes, the present one solves a trivial Bezout equation which does not require any matrix inversion.

III. THE ADAPTIVE CASE: CONSTRUCTION OF THE ERROR MODEL

Let us now consider the ideal control law expressed by (8), (11), (14), and (17). When the plant is unknown, it is possible to choose a control law with the same structure, that is

$$\begin{aligned} u_p(t) &= -\theta_1^T(t)x_1(t) - \theta_2^T(t)x_2(t) - \theta_{20}(t)y_p(t) + w(t) \\ w(t) &= \theta_{30}^T(t)x_3(t) + r(t). \end{aligned} \quad (21)$$

Defining, as usual in the MRAC context,

$$\hat{u}_p(t) = u_p(t) - u_p^*(t) \quad (22)$$

where $u_p^*(t)$ is the ideal control law determined from (8) and (11), one has

$$\begin{aligned} \hat{u}_p(t) &= -(\theta_1^T(t) - \theta_1^{*T})x_1(t) - (\theta_{20}(t) - \theta_{20}^*)y_p(t) \\ &\quad - (\theta_2^T(t) - \theta_2^{*T})x_2(t) + (\theta_{30}^T(t) - \theta_{30}^{*T})x_3(t). \end{aligned} \quad (23)$$

Thus, the auxiliary plant state equation can be rewritten as follows

$$\begin{aligned} \dot{x}_a(t) &= A_a x_a(t) + b_a u_p(t) \\ &= A_a x_a(t) + b_a (u_p^*(t) + \hat{u}_p(t)) \\ y_a(t) &= h_a^T x_a(t) \end{aligned} \quad (24)$$

where $x_a(t) \in \mathcal{R}^{q+1}$ and Λ_a is a $(q+1) \times (q+1)$ matrix. Taking into account (8), the first of (24) becomes

$$\dot{r}_1(t) = \Lambda_a x_a(t) - b_a(\theta_1^{*T} r_1(t) + \theta_2^{*T} r_2(t) + \theta_{20}^* y_a(t)) + b_a(u^*(t) + u_p(t)) \quad (25)$$

This equation is identical to the state equation of the ideal auxiliary plant (i.e., the auxiliary plant in the known parameters case), with an input signal $(u^*(t) + \dot{u}_p(t))$ instead of simply $u^*(t)$. From Theorems 1 and 2 we know that the zero state response of $y_a(t)$ is expressible as the solution of the following first order differential equation

$$\dot{y}_a(t) = -a y_a(t) + (u^*(t) + u_p(t)) \quad (26)$$

Therefore, by introducing a parallel model of the type

$$\begin{aligned} \dot{z}_a(t) &= -a z_a(t) + u(t) \\ z_a(t) &= -a z_a(t) + \theta_1^T(t) r_1(t) + \theta_2^T(t) r_2(t) + \theta_{20}(t) y_a(t), \end{aligned} \quad (27)$$

the final error equation, due to the error on the control signal $u_p(t)$, results

$$\begin{aligned} \nu(t) &= y_1(t) - z_a(t) \\ \dot{\nu}(t) &= -a\nu(t) - \theta_1^T(t) r_1(t) - \theta_2^T(t) r_2(t) - \theta_{20}(t) y_a(t) \\ &= -a\nu(t) + \Theta^T(t) \mathbf{X}(t) \end{aligned} \quad (28)$$

where

$$\begin{aligned} \Theta^T(t) &= [-\theta_1^T(t) - \theta_{20}(t) - \theta_2^T(t)] \\ \mathbf{X}^T(t) &= [r_1^T(t) y_1(t) r_2^T(t)] \end{aligned}$$

Thus the adaptive control problem has been reduced to the standard form, i.e., $\nu(t) = I_m(s)\Theta^T(t)\mathbf{X}(t)$ with $I_m(s) = 1/(s+a) \in \text{SPR}$, and $\nu(t)$ playing the role of tracking error

IV. THE ADAPTIVE CASE: INTRODUCTION OF A COMBINED DIRECT AND INDIRECT ADAPTIVE LAW

Let us consider the ideal control law expressed by (8), and the error model previously derived, i.e. $\nu(t) = -a\nu(t) + \Theta^T(t)\mathbf{X}(t)$ with $\Theta(t) \in \mathcal{R}^m$, $\nu(t) \in \mathcal{R}^1$. We propose the following adjustment law

$$\dot{\Theta}(t) = -\Gamma(t)[\mathbf{X}(t)\nu(t) + a\varphi(t)(\nu(t) - \epsilon(t)) + \varphi(t)\epsilon(t)] \quad (29)$$

where $\varphi(t) \in \mathcal{R}^m$, $\epsilon(t) \in \mathcal{R}^1$, $\Gamma(t)$ is a $m \times m$ matrix, and they are adjusted as follows

$$\dot{\varphi}(t) = -a\varphi(t) + \mathbf{X}(t) \quad (30)$$

$$\begin{aligned} \dot{\epsilon}(t) &= -a\epsilon(t) + \varphi^T(t) \mathbf{I}(t) [\mathbf{X}(t)\nu(t) \\ &\quad + a\varphi(t)(\nu(t) - \epsilon(t)) + \varphi(t)\epsilon(t)] + \epsilon(t)(\nu - \epsilon) \end{aligned} \quad (31)$$

$$\dot{\mathbf{I}}(t) = -a\Gamma(t)\varphi(t)\varphi^T(t)\Gamma(t) + a(\Gamma(t) + \Gamma^T(t)\Gamma(t)/k_0), \quad (32)$$

where $0 < \Gamma(0) \leq k_0 I_m$, $k_0 > 0$ being an upperbound on the gain matrix, and $\epsilon(t)(\nu - \epsilon) \in \mathcal{R}^1$ satisfies the following inequality

$$\epsilon(t)(\nu - \epsilon)(\nu(t) - \epsilon(t)) \geq 0 \quad (33)$$

Note that $(\nu(t) - \epsilon(t))$ is playing the role of the prediction error estimate. The function $\epsilon(\nu, \epsilon)$ can be chosen as $\epsilon(\nu, \epsilon) = \gamma_1 \text{sign}(\nu(t) - \epsilon(t))$, $\gamma_1 > 0$. The introduction of the relay term $\epsilon(\nu, \epsilon)$ driven by the estimate of the prediction error is essential for robustness enhancement, when bounded disturbances are acting on the plant [13].

Moreover, it is worth remarking that the least squares gain update (32) is a special case of the modification of least squares update law proposed by Lozano *et al.* in [17].

By taking into account the combined adaptive algorithm (29)–(32) we can prove the following result

Theorem 3 Consider the error model (28) with adjustment (29)–(32). If the reference input $r(t)$ is bounded then the tracking error $\nu(t)$ and the prediction error estimate $\epsilon(t)$ converge to zero asymptotically and all the trajectories of the system remain bounded. Moreover, if $\mathbf{X}(t)$ is persistently exciting, the tracking error $\nu(t)$ and the parameter error $\Theta(t)$ are globally exponentially convergent.

Proof The stability of the system can be studied by using the following Lyapunov function

$$V = \frac{1}{2}\nu^2(t) + \frac{1}{2}(\nu(t) - \epsilon(t) - \varphi^T(t)\Theta(t))^2 + \frac{1}{2}\|\Theta(t)\|_{\Gamma(t)}^2 \quad (34)$$

whose derivative along the solutions of (28), (29)–(32) is

$$\begin{aligned} \dot{V} &= \nu(-a\nu + \Theta^T\mathbf{X}) + (\nu - \epsilon - \varphi^T\Theta)\{-a\nu \\ &\quad + \mathbf{X}^T\Theta + a\epsilon - \varphi^T\Gamma(t)[\mathbf{X}\nu + a\varphi(\nu - \epsilon) + \varphi\epsilon] - \\ &\quad + a\varphi^T\Theta - \mathbf{X}^T\Theta + \varphi^T\Gamma(t)[\mathbf{X}\nu + a\varphi(\nu - \epsilon) + \varphi\epsilon]\} \\ &\quad - \Theta^T[\mathbf{X}\nu + a\varphi(\nu - \epsilon) + \varphi\epsilon] + \frac{a}{2}\Theta^T\varphi\varphi^T\Theta \\ &\quad - \frac{a}{2}\|\Theta\|_{\Gamma(t)-I_m/k_0}^2 \\ &= -a\nu^2 - a(\nu - \epsilon - \varphi^T\Theta)^2 - (\nu - \epsilon)\epsilon \\ &\quad + \Theta^T\varphi\epsilon - a\Theta^T\varphi(\nu - \epsilon) - \Theta^T\varphi\epsilon \\ &\quad + \frac{a}{2}\Theta^T\varphi\varphi^T\Theta - \frac{a}{2}\|\Theta\|_{\Gamma(t)-I_m/k_0}^2 \end{aligned} \quad (35)$$

(note that for the sake of simplicity the dependence of $\nu, \epsilon, \varphi, \Theta$ on time has been omitted)

Taking into account (33) and decomposing the second term we have

$$\begin{aligned} \dot{V} &\leq -a\nu^2 - \frac{a}{2}(\nu - \epsilon - \varphi^T\Theta)^2 \\ &\quad - \frac{a}{2}(\nu - \epsilon)^2 - \frac{a}{2}\|\Theta\|_{\Gamma(t)-I_m/k_0}^2 \end{aligned} \quad (36)$$

From this we conclude that $\nu(t)$, $(\nu(t) - \epsilon(t) - \varphi^T(t)\Theta(t))$ are bounded and square-integrable, $\Theta(t)$ is bounded and $(\nu(t) - \epsilon(t))$ is square-integrable. Now, let us examine the boundedness of $\dot{\nu}(t)$. Since $r(t)$ is bounded, $r_1(t)$ is bounded as well (see (5)). From (27) we conclude that the boundedness of $r(t)$, $\Theta(t)$ and $r_1(t)$ implies the boundedness of $\dot{z}_a(t)$ and in turn the boundedness of $y_1(t)$ since $\nu(t)$ is bounded. From (5) we conclude that $r_2(t)$ is bounded since y_1 is bounded. Examining the boundedness of $u_1(t)$, from (7) we see that $u_1(t)$ is bounded since the auxiliary plant transfer function has zeros in the open left half plane and according to, for instance, Lemma 4 in [3], its input cannot grow faster than its output, namely $y_1(t)$, which is bounded. From the boundedness of $u_1(t)$ we conclude the boundedness of $r_1(t)$, and that of the output of the parallel compensator $y_1(t)$, as well as, in turn, the boundedness of the plant output $y_1(t)$, since $y_a(t)$ is bounded. From the boundedness of $\mathbf{X}(t)$ it follows the boundedness of $\varphi(t)$ and $\varphi^T(t)$. The boundedness of $\varphi(t)$, $\mathbf{I}(t)$, $\mathbf{X}(t)$, $\nu(t)$ and $\epsilon(t)$ implies the boundedness of $\epsilon(t)$ and $\epsilon(t)$ and hence we conclude the boundedness of $\Theta(t)$. Since $\nu(t)$ and $(\nu(t) - \epsilon(t))$ are square-integrable and $\nu(t)$ and $(\nu(t) - \epsilon(t))$ are bounded, they both converge to zero.

Now consider the case when $\mathbf{X}(t)$ is persistently exciting. The excitation of the regressor $\mathbf{X}(t)$ is connected with the excitation strength of the reference input $r(t)$ [18]. If $\mathbf{X}(t)$ is persistently exciting then $\varphi(t)$ is exciting as well. Note that the modified least squares estimator (32) presents the following properties: a) $\mathbf{I}(t) \leq k_0 I_m$ for any $t \geq 0$, b) if $\varphi(t)$ is persistently exciting, i.e., there exist strictly positive constants δ, β, β_1 such that

$$\beta_1 I_m \geq \int_t^{t+\delta} \varphi(\tau)\varphi^T(\tau) d\tau \geq \beta I_m \quad (37)$$

then, there exist $\lambda_1 > 0$ and $\lambda_2 > 0$ such that $(\Gamma^{-1}(t) - I_m/k_0) \geq \lambda_1 I_m$ and $\Gamma^{-1}(t) \leq \lambda_2 I_m$ for any $t \geq 0$. Indeed, define $z(t) = x^T \Gamma^{-1}(t)x$, where $x \in \mathbb{R}^m$ is any constant unitary vector. Differentiating $z(t)$ and taking into account that $\Gamma^{-1}(t) = -\Gamma^{-1}(t)\dot{\Gamma}(t)\Gamma^{-1}(t)$ we have $\dot{z}(t) = ax^T \varphi(t)\varphi^T(t)x - az(t) + a/k_0 \geq -az(t) + a/k_0$ which implies that $\Gamma(t) \leq k_0 I_m$. Integrating $\dot{z}(t)$ and taking into account (37), we obtain $z(t+h) - z(t) \geq ah - a \int_t^{t+h} z(\tau) d\tau + ah/k_0$. This, in turn implies that there exists $\lambda_1 > 0$ such that $(\Gamma^{-1}(t) - I_m/k_0) \geq \lambda_1 I_m$. Using similar arguments one can show that there exists $\lambda_2 > 0$ such that $\Gamma^{-1}(t) \leq \lambda_2 I_m$ for any $t \geq 0$.

Then, by using the above properties of the estimator (32) we can write (35) as

$$\dot{V} \leq -\frac{a}{2}\nu^2 - \frac{a}{2}(\nu - \epsilon - \varphi^T \dot{\Theta})^2 - \frac{a\lambda_1}{2\lambda_2} \|\dot{\Theta}\|^2. \quad (38)$$

Let a_1 be a positive constant such that $a_1 = \min[a, a\lambda_1/\lambda_2]$. Then $\dot{V} \leq -a_1 V$ which implies the exponential convergence of $\nu(t)$, $(\nu(t) - \epsilon(t) - \varphi^T(t)\dot{\Theta}(t))$ and $\dot{\Theta}(t)$ to zero. \square

To conclude the treatment, it should be noted that no direct adaptation of the parameter vector $\theta_3(t)$ is performed, since this vector does not affect the output error $\nu(t)$. Yet, the parameters $\theta_3(t)$ can be indirectly adjusted on the basis of (20). Since the exponential convergence of $\theta_1(t)$, $\theta_2(t)$, $\theta_{20}(t)$ to their ideal values has been ensured, also the convergence of $\theta_3(t)$ to θ_3^* is guaranteed. This corresponds to the attainment of the control objective (3) in Problem 1.

V. CONCLUSIONS

In this note, an alternative adaptive control approach is presented so as to deal with plants with unknown relative degree and uncertain order which can be made minimum phase by means of a relative-degree-one first order parallel compensator. Apart from the fundamental feature of having a complexity which turns out to be independent of the relative degree of the plant, the proposed control scheme is characterized by the use of a nonstandard adaptation mechanism which includes a time-varying gain and can be viewed as a combined direct/indirect mechanism, being based on the use of a prediction error estimate. The introduction of this modified adaptation mechanism allows us to prove that the tracking error asymptotically converges to zero and that all the trajectories of the system remain bounded. Moreover, under a suitable assumption of persistent excitation, the tracking error and the parameter error turn out to be globally exponentially convergent. This latter property is indirectly used to complete the design of the control scheme so as to assign the desired dynamics to the true plant output.

REFERENCES

- [1] I. D. Landau, *Adaptive Control: The Model Reference Approach*. New York: Dekker, 1979.
- [2] R. V. Monopoli, "Model reference adaptive control with an augmented error signal," *IEEE Trans. Automat. Contr.*, vol. 19, pp. 474-484, 1974.
- [3] K. S. Narendra, Y. H. Lin, and L. S. Valavani, "Stable adaptive controller design, part II: Proof of stability," *IEEE Trans. Automat. Contr.*, vol. 25, pp. 440-448, 1980.
- [4] P. A. Ioannou and P. V. Kokotovic, "Robust redesign of adaptive control," *IEEE Trans. Automat. Contr.*, vol. 29, pp. 202-211, 1984.
- [5] P. A. Ioannou and K. S. Tsakalis, "A robust direct adaptive controller," *IEEE Trans. Automat. Contr.*, vol. 31, pp. 1033-1043, 1986.

- [6] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*. New Jersey: Prentice Hall, 1989.
- [7] S. M. Naik, P. R. Kumar, and B. E. Ydsdtie, "Robust continuous time adaptive control by parameter projection," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 182-197, 1992.
- [8] G. Bartolini, and A. Ferrara, "A new adaptive pole assignment control scheme," *J. Syst. Eng.*, vol. 2, pp. 134-142, 1992.
- [9] —, "Adaptive control of SISO plants with unmodeled dynamics," *Int. J. Adaptive Contr. Signal Proc.*, vol. 6, pp. 237-246, 1992.
- [10] J. J. E. Slotine and W. Li, "Compositional adaptive control of robot manipulators," *Automatica*, vol. 25, pp. 509-519, 1989.
- [11] R. H. Middleton and G. C. Goodwin, "Adaptive computed torque control for rigid link manipulators," *Syst. Contr. Lett.*, vol. 2, pp. 9-16, 1988.
- [12] V. O. Nikiforov, "A Stable gradient algorithm of adaptation using an output signal," *Int. J. Adapt. Contr. Signal Proc.*, vol. 6, pp. 265-269, 1992.
- [13] A. Stotsky, "Combined adaptive and variable structure control: New generalized control schemes," *Proc. IEEE Workshop on VSC*, Sheffield, U.K., pp. 162-165, 1992.
- [14] R. Ortega, G. Bartolini, and A. Ferrara, "On zero relocation of plants with structured uncertainties," in *Proc. IECON'94*, Italy, 1994, pp. 1724-1726.
- [15] G. Bartolini and T. Zolezzi, "The VSS approach to the model reference control of nonminimum phase linear plants," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 859-863, 1988.
- [16] G. Bartolini and A. Ferrara, "Discontinuous control of uncertain non-minimum phase plants" in *Proc. 32nd IEEE CDC*, San Antonio, TX, 1993, pp. 2458, 2464.
- [17] R. Lozano, J. Collado, and S. Mondie, "Model reference robust adaptive control without a priori knowledge of the high frequency gain," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 71-78, 1990.
- [18] S. Boyd and S. S. Sastry, "On parameter convergence in adaptive control," *Syst. Contr. Lett.*, vol. 3, pp. 311-319, 1983.

Preservation of Reachability and Observability under Sampling with a First-Order Hold

Tomomichi Hagiwara

Abstract—This paper gives the necessary and sufficient condition for the reachability of the sampled-data system S_1 obtained by the discretization of a linear time-invariant continuous-time system with a first-order hold. Equivalence of the reachability and controllability of S_1 is also shown. Similar results are given also for observability and reconstructibility. It turns out that S_1 is reachable only if S_0 is reachable, while S_1 is observable if and only if $S_1/0$ is observable, where S_0 is the sampled-data system obtained by the discretization with a zero-order hold of the same sampling period.

I. INTRODUCTION

In sampled-data control, hold circuits are used to convert the discrete-time signals from digital compensators into the continuous-time signals to be applied to the continuous-time systems. Hold circuits can be viewed also as filters which attenuate the high-frequency alias spectra generated by sampling continuous-time signals. Typical hold circuits are a zero-order hold and a first-order hold [1], but the former seems to be particularly popular in industrial applications. The primary reason for this is that a zero-order hold can be implemented

Manuscript received October 20, 1993; revised February 20, 1994.
The author is with the Department of Electrical Engineering, Kyoto University, Yoshida, Sakyo-ku, Kyoto 606-01, Japan.
IEEE Log Number 9407018.

quite easily by using the function of D/A converters while a first-order hold can be implemented only with the aid of some additional analog circuits. Another reason might be that, when viewed as continuous-time filters, the phase lag of a first-order hold is greater than that of a zero-order hold for high-frequency ranges, which seems to be a disadvantage from the point of view of closed-loop stability. However, the latter reason seems to apply mainly in the case when a digital compensator is obtained by a digital redesign method [2] of a continuous-time compensator and closed-loop stability is not necessarily assured theoretically. If we could use a first-order hold in such a way that closed-loop stability can be assured, then it might provide some advantages over a zero-order hold, such as reduction of the intersample ripple of the response.

Based upon the above consideration, the aim of this paper is to provide a basis for the use of a first-order hold in the context of the state-space approach of control system design. For this purpose, we give the necessary and sufficient condition for the reachability of the sampled-data system obtained by the discretization of a linear time-invariant continuous-time system with a first-order hold. In addition, we show the equivalence of the reachability and controllability of this sampled-data system. Furthermore, we give similar results for observability and reconstructibility. (For the standard definitions of these concepts, see [3].)

II. DISCRETIZATION WITH A FIRST-ORDER HOLD

We consider the system given by

$$\frac{dx}{dt} = Ax + Bu, \quad y = Cx, \quad (1)$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, and $C \in \mathbb{R}^{p \times n}$. Suppose a first-order hold is connected to the input. Then, $u(t)$ is given by

$$u(t) = u(kT) + \frac{u(kT) - u(\bar{k} - 1T)}{T}(t - kT) \quad (kT \leq t < \bar{k} + 1T), \quad (2)$$

where T denotes the sampling period ($u(kT)$ stands for $u(kT + 0)$). It should be noted that there is a built-in constraint that the input $u(t)$ ($kT \leq t < \bar{k} + 1T$) depends not only on $u(kT)$ but also on $u(\bar{k} - 1T)$, which shows sharp contrast with a zero-order hold. In particular, $u(t)$ ($0 \leq t < T$) depends on $u(-T)$, which has been determined before $t = 0$ and cannot be changed by the compensator at $t = 0$.

The resulting sampled-data system can be described by the equation (see [4], [5])

$$\begin{aligned} x(\bar{k} + 1T) &= Ax(kT) + B^+ u(kT) + B^- u(\bar{k} - 1T), \\ y(kT) &= Cx(kT) \end{aligned} \quad (3)$$

where,

$$\begin{aligned} A &= \exp(AT), \quad B^+ = \int_0^T \left(2 - \frac{t}{T}\right) \exp(A, t) B, dt, \\ B^- &= \int_0^T \left(\frac{t}{T} - 1\right) \exp(A, t) B, dt. \end{aligned} \quad (4)$$

We denote the system (3) by S_1 , which can be rewritten in the form of the ordinary discrete-time state equation as

$$\begin{bmatrix} x(\bar{k} + 1T) \\ u(kT) \end{bmatrix} = \begin{bmatrix} A & B^- \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(kT) \\ u(\bar{k} - 1T) \end{bmatrix} + \begin{bmatrix} B^+ \\ I_m \end{bmatrix} u(kT). \quad (5)$$

$$y(kT) = [C \ 0] \begin{bmatrix} x(kT) \\ u(\bar{k} - 1T) \end{bmatrix}. \quad (6)$$

III. CONTROLLABILITY AND REACHABILITY OF S_1

A. Definitions of Controllability and Reachability and Their Equivalence

In this subsection, we first give the definitions of the controllability and reachability of S_1 . In view of the discrete-time state equation (5), let us adopt the following definition.

Definition 1: S_1 is controllable if the pair (A_1, B_1) is controllable, where

$$(A_1, B_1) := \left(\begin{bmatrix} A & B^- \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} B^+ \\ I_m \end{bmatrix} \right)$$

Now, let us verify that the above formal definition matches our practical control purposes in spite of the built-in constraint of a first-order hold.

If we regard S_1 simply as an ordinary discrete-time system, then its controllability might be defined as the property that, given any initial condition $x(0)$, there exists a sequence $u(kT)$ ($k = 0, \dots, N-1$) such that $x(NT) = 0$. However, this is not appropriate, because this property does not reflect real purposes of control. Namely, this definition does not always imply the property that there exists a sequence $u(kT)$ ($k = 0, \dots, N-1, N, \dots$) such that $x(t) = 0$ ($\forall t \geq NT$), because of the built-in constraint of a first-order hold as discussed in the previous section. (This can be easily understood if we notice that $u(t) = 0$ ($NT \leq t < \bar{N} + 1T$) implies $u(\bar{N} - 1T) = u(NT) = 0$ from (2).) Therefore, to define controllability, we must require that there exists a sequence $u(kT)$ ($k = 0, \dots, N-2$) such that this together with $u(\bar{N} - 1T) = 0$ implies $x(NT) = 0$. Likewise, as discussed in the previous section, $u(t)$ ($0 \leq t < T$) is constrained by the unprescribable value $u(-T)$. Despite this constraint, $x(NT)$ is required to be made 0.

From the above consideration, the controllability of S_1 should be defined as the property that, given any initial conditions $x(0)$ and $u(-T)$, there exists a sequence $u(kT)$ ($k = 0, \dots, N-2$) such that this together with $u(\bar{N} - 1T) = 0$ implies $x(NT) = 0$. Obviously, this definition is equivalent to the controllability of the discrete-time system (5). Thus, validity of the above definition is assured.

Similarly, we are led to the following definition (see [6] for details).

Definition 2: S_1 is reachable if the pair (A_1, B_1) is reachable.

Now, in spite of the singularity of A_1 , we can establish the following result (the straightforward proof [6] is omitted here).

Theorem 1. S_1 is reachable if and only if it is controllable.

B. Condition for Preservation of Reachability

In this subsection, we study the necessary and sufficient condition for the reachability of S_1 in terms of A , B , and T . From Definition 2, it is reachable if and only if

$$\text{rank} \begin{bmatrix} A - zI_n & B & B^+ \\ 0 & I_m - zI_m & I_m \end{bmatrix} = n + m \quad (\forall z \in \mathbb{C}) \quad (8)$$

where $B := B^+ + B^- = \int_0^T \exp(A, t) B, dt$. (Note that (A, B) is nothing but the pair of the sampled-data system obtained by the discretization of (1) with a zero-order hold, which we denote by S_0 .) The condition (8) is nothing but the reachability condition for the pair

$$(A_2, B_2) := \left(\begin{bmatrix} A & B \\ 0 & I_m \end{bmatrix}, \begin{bmatrix} B^+ \\ I_m \end{bmatrix} \right). \quad (9)$$

This pair can be regarded as the pair obtained by the discretization of the fictitious T -dependent continuous-time pair

$$(A_2, B_2) := \left(\begin{bmatrix} A & B \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} B \\ I_m/T \end{bmatrix} \right) \quad (10)$$

with a zero-order hold, because $A_2 = \exp(A, T)$, $B_2 = \int_0^T \exp(A_2, t) B_2, dt$.

Remark 1: (A_{2t}, B_{2t}) is reachable if and only if a) (A_t, B_t) is reachable and b) A_t does not have the eigenvalue $-1/T$.

Since the eigenvalues and left eigenvectors of A_{2t} are not dependent on T , we can apply the necessary and sufficient condition [7] for the reachability of S_0 to the pair (A_{2t}, B_{2t}) . Then, the following theorem is obtained (see Appendix for proof).

Theorem 2: S_1 is reachable if and only if the following two conditions are satisfied.

- a) S_0 is reachable.
- b) A_t does not have the eigenvalue $-1/T$.

Remark 2: Suppose that we define the stabilizability of S_1 by the stabilizability of the pair (A_1, B_1) . Then, the necessary and sufficient condition for the stabilizability of S_1 is given by the stabilizability of S_0 (the condition b) can be dropped). Although we can rewrite the reachability/stabilizability of S_0 as the conditions on A_t, B_t , and T (see [7], and Theorem A.1 in Appendix), we did not do this because the importance of the theorem seems to be much clearer in the present form of the statement.

IV. OBSERVABILITY AND RECONSTRUCTIBILITY OF S_1

As in the preceding section, let us consider how to define the observability and reconstructibility of S_1 , taking account of practical purposes.

If we regard S_1 simply as an ordinary discrete-time system, then its observability might be defined as the property that its initial state $x(0)$ can be uniquely determined from the input data $u(kT)$ ($k = 0, \dots, N-1$) and the output data $y(kT)$ ($k = 0, \dots, N$). However, this is not appropriate, because $u(t)$ ($0 \leq t < T$) cannot be known completely from the knowledge of the above input data, as discussed in Section II, and it is clearly impossible to determine $x(0)$ under this lack of knowledge. Therefore, to define observability, we must assume that $u(t)$ ($0 \leq t < T$) is also known. This assumption is equivalent to the assumption that $u(-T)$ as well as the above input and output data can be used. Noting that $x(0)$ can be determined uniquely if and only if $[x(0)^T, u(-T)^T]^T$ can be determined uniquely (if we know $u(-T)$), we are led to the following definition.

Definition 3: S_1 is observable if the pair (C_1, A_1) is observable, where

$$C_1 := \begin{bmatrix} C' & 0 \\ 0 & I_m \end{bmatrix}. \quad (11)$$

Similarly, we are led to the following definition (see [6] for details).

Definition 4: S_1 is reconstructible if the pair (C_1, A_1) is reconstructible, where

$$C_1 := [C' \ 0]. \quad (12)$$

(C_1, A_1) is reconstructible if and only if (C_1, A_1) is reconstructible. Furthermore, in spite of the singularity of A_1 , we can readily show that (C_1, A_1) is reconstructible if and only if it is observable. Thus we obtain the following result.

Theorem 3: S_1 is observable if and only if it is reconstructible.

Now, from Definition 3, S_1 is observable if and only if

$$\text{rank} \begin{bmatrix} A - zI_n & B^- \\ 0 & -zI_m \\ C' & 0 \\ 0 & I_m \end{bmatrix} = n + m \quad (\forall z \in \mathbb{C}). \quad (13)$$

Since (C', A) is the pair of S_0 , we readily obtain the following theorem.

Theorem 4: S_1 is observable if and only if S_0 is observable.

Remark 3: Suppose that we define the detectability of S_1 by the detectability of the pair (C_1, A_1) . Then, the necessary and sufficient condition for the detectability of S_1 is given by the detectability of S_0 . The condition for observability/detectability of S_0 in terms of C, A_t and T is given by [7] (see also Remark A.1 in Appendix).

V. CONCLUSION

In this paper, we studied the use of a first-order hold in the context of the state-space approach of control system design. We first studied how to define the controllability and reachability for the sampled-data system S_1 obtained by the discretization of a linear time-invariant continuous-time system with a first-order hold, taking account of the built-in constraint of a first order hold. Next, we showed the equivalence of these two concepts for S_1 . Then, we studied the necessary and sufficient condition for the reachability of S_1 in terms of the parameters of the continuous-time system and the sampling period. We also gave similar results for observability and reconstructibility. In particular, it turned out that S_1 is reachable only if S_0 is reachable, while S_1 is observable if and only if S_0 is observable, where S_0 is the sampled-data system for the zero-order hold case. The compensator design problem under the use of a first-order hold is also studied in [6].

APPENDIX

PROOF OF THEOREM 2

Before proving Theorem 2, we give a more comprehensible statement of the necessary and sufficient condition for the reachability of S_0 derived in [7].

Let $\lambda(A_t)$ denote the set of the eigenvalues of A_t . For each $\lambda_i \in \lambda(A_t)$, we define

$$\Lambda(\lambda_i) := \{\lambda \mid \lambda \in \lambda(A_t), \text{Re}(\lambda) = \text{Re}(\lambda_i), \text{Im}(\lambda) - \text{Im}(\lambda_i) = 2k\pi/T \quad (k = 0, \pm 1, \pm 2, \dots)\}. \quad (14)$$

Note that $\lambda_i \in \Lambda(\lambda_j)$, and that $\Lambda(\lambda_i) = \Lambda(\lambda_j)$ if $\lambda_i \in \Lambda(\lambda_j)$. Our interest is only in the sets $\Lambda(\lambda_i)$ which have at least two elements. Let Λ_l ($l = 1, \dots, L$) be such distinct sets, where we assume that Λ_l ($l = 1, \dots, L^+ (\leq L)$) are the sets corresponding to the eigenvalues with nonnegative real parts (L and L^+ might be zero). We denote the elements of Λ_l by λ_{lk} ($k = 1, \dots, K_l$).

Next, for each $\lambda_i \in \lambda(A_t)$, we define

$$\Gamma(\lambda_i) := \begin{bmatrix} \eta_{i,1}^T \\ \vdots \\ \eta_{i,\nu_i}^T \end{bmatrix}, \quad (15)$$

where ν_i denotes the geometric multiplicity of the eigenvalue λ_i , and $\eta_{i,k}^T$ ($k = 1, \dots, \nu_i$) the corresponding linearly independent left eigenvectors. That is to say, all the linearly independent left eigenvectors of A_t corresponding to the eigenvalue λ_i form the rows of $\Gamma(\lambda_i)$. We further define

$$\Gamma(\Lambda_l) := \begin{bmatrix} \Gamma(\lambda_{l1}) \\ \vdots \\ \Gamma(\lambda_{lK_l}) \end{bmatrix} \quad (16)$$

for $l = 1, \dots, L$. That is to say, all the linearly independent left eigenvectors of A_t corresponding to the eigenvalues in the set Λ_l form the rows of $\Gamma(\Lambda_l)$.

Now, we obtain the following theorem, which is merely a restatement of Theorem 2 of [7].

Theorem A.1: S_0 is reachable (respectively, stabilizable) if and only if the following three conditions hold:

- a) (A_l, B_l) is reachable (respectively, stabilizable).
- b) A_l does not have the nonzero eigenvalue

$$j2k\pi/T \quad (k = \pm 1, \pm 2, \dots). \quad (17)$$

- c) $\Gamma(\Lambda_l)B_l$ has full row rank for $l = 1, \dots, L$ (respectively, $l = 1, \dots, L^+$).

Remark A.1: The conditions for the observability and the detectability of S_0 are given by the dual of the conditions a) and c) of the above theorem. (This is a restatement of Theorem 3 of [7].)

Remark A.2: Note that $\Gamma(\Lambda_l)$ never contains the left eigenvectors for the zero eigenvalue of A_l if the condition b) holds.

Now, we give the proof of Theorem 2.

Proof of Theorem 2: Without loss of generality, we assume that

$$A_l = \begin{bmatrix} \hat{A}_l & 0 \\ 0 & Z \end{bmatrix}, \quad B_l = \begin{bmatrix} \hat{B}_{l1} \\ \hat{B}_{l2} \end{bmatrix}, \quad (18)$$

where \hat{A}_l is nonsingular and all the eigenvalues of Z are zero. Then, applying the similarity transformation by the matrix

$$\begin{bmatrix} I & 0 & -A_l^{-1}\hat{B}_{l1} \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \quad (19)$$

to the pair (10), we obtain the pair

$$\left(\begin{bmatrix} A_l & 0 & 0 \\ 0 & Z & B_{l2} \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} (I + \hat{A}_l^{-1}/T)\hat{B}_{l1} \\ B_{l2} \\ I/T \end{bmatrix} \right). \quad (20)$$

Applying Theorem A.1 to the pair (10) for the conditions a) and b), and to the pair (20) for the condition c), and taking Remarks 1 and A.2 into account, we can obtain the following necessary and sufficient condition for the reachability of S_1 :

- A1) (A_l, B_l) is reachable.
- A2) A_l does not have the eigenvalue $-1/T$.
- B) A_l does not have the nonzero eigenvalue

$$j2k\pi/T \quad (k = \pm 1, \pm 2, \dots). \quad (21)$$

- C) $\Gamma(\hat{\Lambda}_l)(I + A_l^{-1}/T)\hat{B}_{l1}$ has full row rank for $l = 1, \dots, \hat{L}$, where $\hat{\Lambda}_l$, $\Gamma(\hat{\Lambda}_l)$ and \hat{L} are defined for \hat{A}_l in a consistent way with the above definitions of Λ_l , $\Gamma(\Lambda_l)$ and L .

Since the rows of $\hat{\Gamma}(\Lambda_l)$ are the left eigenvectors of \hat{A}_l by definition, under condition A2) the condition C) is equivalent to

$$C') \hat{\Gamma}(\Lambda_l)\hat{B}_{l1} \text{ has full row rank for } l = 1, \dots, \hat{L}.$$

In view of the form of (18), the conditions B) and C') are equivalent to the conditions b) and c) of Theorem A.1. Since the condition A1) is the same as the condition a) of Theorem A.1, and since the condition (A2) is the same as the condition b) of Theorem 2, the proof has become complete. Q.E.D.

ACKNOWLEDGMENT

The author wishes to express his gratitude to Prof. M. Kimura, motivating the research through discussions.

REFERENCES

- [1] J. T. Tou, *Digital & Sampled-Data Control Systems*. New York: McGraw-Hill, 1959.
- [2] P. Katz, *Digital Control Using Microprocessors*. Englewood Cliffs, NJ: Prentice-Hall, 1981.
- [3] R. E. Kalman, P. L. Falb, and M. A. Arbib, *Topics in Mathematical System Theory*. New York: McGraw-Hill, 1969.
- [4] T. Mita, *Theory of Digital Control*. Tokyo: Shoukou Do, 1984 (in Japanese).
- [5] T. Hagiwara, T. Yuasa, and M. Araki, "Stability of the limiting zeros of sampled-data systems with zero- and first-order holds," *Int. J. Control*, vol. 58, no. 6, pp. 1325-1346, 1993.
- [6] T. Hagiwara, "Preservation of reachability and observability under sampling with a first-order hold," Technical Report No. 93-10, Automatic Control Engineering Group, Department of Electrical Engineering II, Kyoto University, 1993.
- [7] M. Kimura, "Preservation of stabilizability of a continuous time-invariant linear system after discretization," *Int. J. Syst. Sci.*, vol. 21, no. 1, pp. 65-91, 1990.

On Robust Asymptotic Tracking: Perturbations on Coprime Factors and Parameterization of All Solutions

Gilberto O. Corrêa and Marcos A. da Silva

Abstract—The robust asymptotic tracking problem is analyzed in this paper relative to unstructured perturbations on each coprime factor of the transfer functions from the control input to the measured and controlled outputs. In each case, necessary and sufficient conditions for the existence of solutions are presented which are explicitly given in terms of problem data. Under such conditions, explicit parameterizations are given of all controllers which achieve robust asymptotic tracking, in terms of free, rational proper, and stable matrices.

I. NOMENCLATURE

\mathbb{C}	Set of complex numbers.
\mathbb{R}_p	Set of real-rational and proper functions from \mathbb{C} to \mathbb{C} .
\mathbb{S}	Ring of real-rational, stable, and proper functions from \mathbb{C} to \mathbb{C} .
$\mathcal{M}(X)$	Set of matrices with entries in X .
$\text{GLCD}\{A, B\}$	Greatest left common divisor of A and B .
A	Column vector obtained by stacking the rows of A .
$A \otimes B$	Kronecker product of A and B .
RCP	Right coprime.
LCP	Left coprime.
ESP	Externally skewprime.

Manuscript received June 17, 1993; revised November 29, 1993 and March 23, 1994.

G. O. Corrêa is with Laboratório Nacional de Computação Científica—LNCC, C.P. 56018, 22290, Rio de Janeiro, RJ, Argentina.

M. A. da Silva is with the Departamento de Engenharia Elétrica—PUCRJ, C.P. 38063, 22453, Rio de Janeiro, RJ, Argentina.

IEEE Log Number 9407019.

II. INTRODUCTION

The robust asymptotic tracking problem (RATP) can be roughly stated as that of finding a stabilizing controller which ensures asymptotic tracking of a given class of reference signals for any plant model P "sufficiently close" to a nominal model P_0 . Early references on this problem include [1]–[4]. The emergence of factorization methods led to fresh approaches to this problem, among which one could mention [5]–[7]. In contrast with previous work which considered unity feedback systems, [8]–[10] addressed the RATP for plant-sensor systems and two-degree-of-freedom controllers. In particular, [10] considered perturbations in the plant model (and not in the sensor) and presented a parameterization of the set of solutions where the "free" parameter must satisfy a "skew" diophantine equation.

The RATP for four-block multivariable linear systems (following [11]) was considered in [12] and [13]. The former considered the robust asymptotic tracking problem relative to perturbations only in the numerator of the transfer function relating the control u to the measured output z and, separately, relative to perturbations only in the denominator of the same transfer function—essentially, the existence of solutions to the RATP relative to perturbations on each coprime factor is shown to be equivalent to the existence of a stabilizing controller which satisfies an internal model condition on its corresponding coprime factor. The latter considered the RATP relative to perturbations constrained by a special linear relation involving the numerators of the transfer functions from u to the controlled and measured outputs (y and z , respectively).

The present note considers perturbed model classes obtained by allowing a single coprime factor of the transfer functions from u to y and z to vary while the others are held fixed, with the following aims: to give explicit necessary and sufficient conditions on problem data for the existence of solutions to the RATP relative to such classes; to give an explicit parameterization of all controllers which solve the RATP relative to such model perturbation classes.

This note is organized as follows. In Section II, the RATP is precisely formulated and the required background material is concisely described. In Sections III, IV, and V, the RATP is considered for perturbed model classes where, respectively, only the numerator of the transfer function from u to z , its denominator, and the numerator or denominator of the transfer function from u to y are allowed to vary. The Lemmas required in the main text can be found in the Appendix; proofs omitted here are to be found in [14]. Notation is essentially the same as in [7].

III. PROBLEM STATEMENT

Consider linear closed-loop systems defined by

$$\begin{bmatrix} y \\ z \end{bmatrix} = P \begin{bmatrix} u \\ d \end{bmatrix} = \begin{bmatrix} P_1 & \bar{P}_1 \\ P_2 & \bar{P}_2 \end{bmatrix} \begin{bmatrix} u \\ d \end{bmatrix},$$

$$\dot{u} = C \begin{bmatrix} r \\ \dot{z} \end{bmatrix} = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} r \\ \dot{z} \end{bmatrix}$$

$$u = \hat{u} + w, \quad z = \hat{z} + v,$$

where P and C are, respectively, the plant and controller transfer functions, $P_i, \bar{P}_i, C_i, i = 1, 2$, are real-rational and proper matrices, d, w, v denote disturbances, r the reference signal, u the control variable, y the controlled output, and z the measurement output. The closed-loop transfer matrix from r to y is $P_1 Q_1$, where $Q_1 \triangleq (I - C_2 P_2)^{-1} C_1$.

Let a class of reference signals be given by $S_r = \{r = \psi_r^{-1} \mu_r : \mu_r \in \mathcal{M}(\mathcal{S}) \text{ and is strictly proper}\}$, where ψ_r is a given biproper, square, and nonsingular matrix in $\mathcal{M}(\mathcal{S})$ such that the zeroes of its determinant have nonnegative real parts.

Given P and ψ_r , the so-called asymptotic tracking problem is to find a stabilizing controller C such that the error $e = r - y = (I - P_1 Q_1)r = (I - P_1 Q_1)\psi_r^{-1} \mu_r$ goes to zero as time goes to infinity for all strictly proper μ_r in $\mathcal{M}(\mathcal{S})$ (hence, in particular, for all strictly proper μ_r in $\mathcal{M}(\mathcal{S})$ such that μ_r and ψ_r are left coprime). By Lemma 5.7.6 [7], this is achieved iff $(I - P_1 Q_1)\psi_r^{-1} \in \mathcal{M}(\mathcal{S})$.

Accordingly, given P_0 and ψ_r , the so-called RATP can be roughly stated as that of finding a controller that solves the asymptotic tracking problem for P and ψ_r , for all P "sufficiently close" to P_0 . Before introducing a precise definition for the RATP for P_0 and ψ_r , and since, in any case, it involves closed-loop stabilization, the results of the latter problem which are required below are collected in the following theorem (see [11]).

Theorem 1: If $P \in \mathcal{M}(\mathbb{R}_p)$ is stabilizable then:

- for $P_1 = N_1 D_1^{-1}, P_2 = N_2 D_2^{-1}$ RCP factorizations, there exists $M \in \mathcal{M}(\mathcal{S})$ such that $D_2 = D_1 M$.
- The set of all stabilizing controllers is given by:

$$S_C = \{C = (K \hat{N}_2 + Y)^{-1} [R : K \hat{D}_2 - X] : R, K \in \mathcal{M}(\mathcal{S}), \det \{(K \hat{N}_2 + Y)(\infty)\} \neq 0\}$$

where $X, Y \in \mathcal{M}(\mathcal{S})$ are such that $X N_2 + Y D_2 = I, P_2 = D_2^{-1} N_2, \hat{N}_2$ and \hat{D}_2 LCP. The corresponding set of closed-loop transfer matrices from r to (y, z) is $\{[N_1^T : N_2^T]^T R : R \in \mathcal{M}(\mathcal{S})\}$, where $\hat{N}_1 \triangleq N_1 M$.

A natural way of considering plant models close to the nominal one (P_{10}, P_{20}) would be to consider small perturbations on the RCP factors $(N_{20}, D_{20}), (N_{10}, D_{10})$ of $P_{10} = N_{10} D_{10}^{-1}, P_{20} = N_{20} D_{20}^{-1}$. However, condition (I.1) of Theorem 1 implies that

$$P_{10} = N_{10} D_{10}^{-1} = (N_{10} M) D_{20}^{-1} = N_{10} D_{20}^{-1}.$$

As a result, one could consider perturbations on (N_{20}, D_{20}, N_{10}) as the general setup where the RATP would be precisely formulated.

To this effect, define, for $\epsilon > 0$

$$\mathcal{V}(\epsilon, P_0) = \{(P_1, P_2) \in \mathcal{M}(\mathbb{R}_p)^2 : P_2 = N_2 D_2^{-1}, P_1 = \hat{N}_1 D_2^{-1}, \text{ for some } \hat{N}_1, N_2, D_2\}$$

such that N_2, D_2 are RCP and $\|[(N_2 - N_{20})^T, (D_2 - D_{20})^T, (\hat{N}_1 - \hat{N}_{10})^T]^T\|_\infty < \epsilon\}$.

The RATP for a given pair (ψ_r, P_0) relative to a model class which contains (P_{10}, P_{20}) can now be precisely defined.

Definition 1: Let a class of models $\mathcal{V} \subset \mathcal{M}(\mathbb{R}_p)$ be given where $(P_{10}, P_{20}) \in \mathcal{V}$. A controller $C \in \mathcal{M}(\mathbb{R}_p)$ is said to solve the RATP for (ψ_r, P_{10}, P_{20}) relative to \mathcal{V} if and only if there exists $\epsilon_0 > 0$ such that, for any (P_1, P_2) in $\mathcal{V} \cap \mathcal{V}(\epsilon_0, P_0)$, C stabilizes (P_1, P_2) and $(I - P_1 Q_1)\psi_r^{-1} \in \mathcal{M}(\mathcal{S})$.

Consider a doubly coprime factorization $(N_{20}, D_{20}, \hat{N}_{20}, \hat{D}_{20}, N_{10}, Y_0, \hat{N}_{10}, \hat{Y}_{10})$ of P_{20} (see [7, p. 48]), i.e., $P_{20} = N_{20} D_{20}^{-1} =$

$D_{20}^{-1}\dot{N}_{20}$ with

$$\begin{bmatrix} -X_0 & Y_0 \\ \dot{D}_{20} & \dot{N}_{20} \end{bmatrix} \begin{bmatrix} -N_{20} & \dot{X}_0 \\ D_{20} & \dot{Y}_0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} -N_{20} & \dot{X}_0 \\ D_{20} & \dot{Y}_0 \end{bmatrix} \begin{bmatrix} -X_0 & Y_0 \\ \dot{D}_{20} & \dot{N}_{20} \end{bmatrix}.$$

The following proposition gives necessary and sufficient conditions for a controller C to achieve robust asymptotic tracking.

Proposition 1: Let (ψ, P_{10}, P_{20}) be given, $P_{20} = N_{20}D_{20}^{-1}$ RCP, $P_{10} = \dot{N}_{10}D_{20}^{-1}$. Let $\mathcal{V} \subset \mathcal{M}(\mathbb{R}_p)$ be such that $(P_{10}, P_{20}) \in \mathcal{V}$. A controller solves the RATP for (ψ, P_{10}, P_{20}) relative to \mathcal{V} iff C is given by $C(R, K) = \dot{D}_C^{-1}(K)[R\dot{N}_C(K)]$, $\dot{N}_C(K) = K\dot{D}_{20} - X_0$, $\dot{D}_C(K) = K\dot{N}_{20} + Y_0$, for some R, K in $\mathcal{M}(\mathcal{S})$ such that $\det\{(K\dot{N}_2 + Y_0)(\infty)\} \neq 0$ and

- $(I - \dot{N}_{10}R)\psi^{-1} \in \mathcal{M}(\mathcal{S})$;
- there exists $\epsilon_0 > 0$ such that for any $(P_1, P_2) \in \mathcal{V} \cap \mathcal{V}(\epsilon_0, P_0)$

$$\begin{aligned} &[\Delta\dot{N}_1 + \dot{N}_{10}(\dot{N}_C(K)\Delta N_2 - \dot{D}_C(K)\Delta D_2)] \\ &\cdot (I - N_C(K)\Delta N_2 + \dot{D}_C(K)\Delta D_2)^{-1}\psi^{-1} \in \mathcal{M}(\mathcal{S}), \end{aligned}$$

where $R\psi^{-1} = \tilde{\psi}^{-1}R$ are coprime factorizations, $\dot{N}_1 = \dot{N}_{10} + \Delta\dot{N}_1$, $N_2 = N_{20} + \Delta N_2$, $D_2 = D_{20} + \Delta D_2$.

Remark: Note that condition a) corresponds to the requirement of nominal asymptotic tracking and that b) is obtained by taking the difference between the perturbed and the nominal $(I - P_1Q_1)\psi^{-1}$.

IV. NUMRATOR PERTURBATIONS ON P_2

In this section the RATP for (ψ, P_{10}, P_{20}) is addressed relatively to the model class \mathcal{N} in which N_2 can vary freely around N_{20} while N_1 and D_2 are held fixed, i.e.,

$$\mathcal{N} \triangleq \{(\dot{N}_{10}D_{20}^{-1}, N_2D_{20}^{-1}) : N_2 \in \mathcal{M}(\mathcal{S}), N_2 \text{ and } D_{20} \text{ RCP}\}.$$

To this effect, note first that $C(R, K)$ solves the RATP for (ψ, P_{10}, P_{20}) relative to \mathcal{N} if and only if R satisfies condition a) and $\dot{N}_{10}\dot{N}_C(K)\Delta N_2(I - \dot{N}_C(K)\Delta N_2)^{-1}\psi^{-1} \in \mathcal{M}(\mathcal{S})$ for all ΔN_2 such that $\|\Delta N_2\|_\infty < \epsilon$ for some $\epsilon > 0$ (this follows from Proposition 1). The aforementioned condition is equivalent to $N_{10}\dot{N}_C(K)\varphi^{-1} = \dot{N}_{10}(K\dot{D}_{20} - X_0)\varphi^{-1} \in \mathcal{M}(\mathcal{S})$, due to Lemma 1, where φ is the greatest invariant factor of ψ , hence the greatest invariant factor of $\tilde{\psi}$ (because $R\psi$ are coprime factorizations). This condition, in turn, is equivalent to $N_\varphi D_\varphi^{-1}\dot{N}_C(K) \in \mathcal{M}(\mathcal{S})$, where $\dot{N}_{10}\varphi^{-1} = N_\varphi D_\varphi^{-1}$, N_φ and D_φ RCP. Due to [7, Lemma 5.7.6], the aforementioned condition is equivalent to $D_\varphi^{-1}\dot{N}_C(K) \in \mathcal{M}(\mathcal{S})$, i.e.,

$$\text{there exists } \bar{N}_C \in \mathcal{M}(\mathcal{S}) \text{ such that } -D_\varphi\bar{N}_C + K\dot{D}_{20} = X_0. \quad (1)$$

Now, $K\dot{D}_{20} - D_\varphi\bar{N}_C = X_0$ implies $K\dot{N}_{20}D_{20} - D_\varphi\bar{N}_C N_{20} = X_0 N_{20}$ and hence (since $X_0 N_{20} + Y_0 D_{20} = I$) $(K\dot{N}_{20} + Y_0)D_{20} + D_\varphi(-\bar{N}_C N_{20}) = I$. Thus, Condition 1 implies that D_φ and D_{20} are ESP. Conversely, suppose that there exists $(\dot{X}, \dot{Y}) \in \mathcal{M}(\mathcal{S}) \times \mathcal{M}(\mathcal{S})$ such that $D_\varphi\dot{X} + \dot{Y}D_{20} = I$. Then, $D_\varphi\dot{X}X_0 + \dot{Y}D_{20}X_0 = X_0$, which is equivalent to $D_\varphi(\dot{X}X_0) + (\dot{Y}Y_0)\dot{D}_{20} = X_0$ (since $D_{20}X_0 = Y_0\dot{D}_{20}$). Thus, if D_φ and D_{20} are ESP, (1) is satisfied by $K = \dot{Y}Y_0$ and $\bar{N}_C = -\dot{X}X_0$. Finally, whenever D_φ and D_{20} are ESP, the set of K which satisfy Condition 1 is obtained by directly applying Lemma 3 to (1). The following theorem can then be stated.

Theorem 2: Let (ψ, P_{10}, P_{20}) be as above

- The set C_1 of all controllers which solve the RATP for (ψ, P_{10}, P_{20}) relative to \mathcal{N} is nonempty iff \mathcal{N} and ψ are ESP and D_φ and D_{20} are ESP.
- In this case, C_1 is given by

$$C_1 = \{(K\dot{N}_{20} + Y_0)^{-1}[R(K\dot{D}_{20} - X_0)] : (R, K) \in S_1, \det\{(K\dot{N}_{20} + Y_0)(\infty)\} \neq 0\}$$

where $S_1 = \{(R, K) : R = \dot{X}\dot{I} - D, R, K = -Y_0\dot{X}_0 - \dot{D}_\varphi\dot{K}, (\dot{R}, \dot{K}) \in \mathcal{M}(\mathcal{S}) \times \mathcal{M}(\mathcal{S})\}$;

$$M_1 = \text{GLCD}\{(\dot{N}_{10} + I), (I - \psi^{-1}), (N_{10} - I) = M_1 N_{10}, (I \times \psi^{-1}) = M_1 \psi^{-1},$$

\dot{I}, \dot{X} and D_φ are defined by $I = M_1 \dot{I}, N_{10}\dot{X} + \psi^{-1}Y = I$, and $\psi^{-1}\dot{N}_{10} = N$, D_φ^{-1} (an RCP factorization), the existence of $\dot{I}, \dot{X}, \dot{Y}, N$, and D_φ being a consequence of the externally skew-primeness of \dot{N}_{10} and ψ ;

$$M_N = \{\text{GLCD}\{(D_\varphi^{-1} - I), (I - \dot{D}_{20}^I)\}, (D_\varphi^{-1} - I) = M_N D_\varphi^{-1}, (I - \dot{D}_{20}^I) = M_N \dot{D}_{20}, \dot{X}_0, Y_0, D_\varphi$$

are defined by $X_0 = M_N \dot{X}_0, D_\varphi Y_0 + \dot{D}_{20}Y_0 = I$, and $\dot{D}_{20}\dot{D}_\varphi = \dot{D}_\varphi\dot{D}_{20}$ which exist since D_φ and D_{20} are ESP.

Proof: In the light of the derivation preceding Theorem 2, only the expression for S_1 needs a proof. For that, note that Lemma 3 applied to (1) and to $\dot{N}_{10}R + W\psi^{-1} = I$ gives the desired expression.

V. DENOMINATOR PERTURBATION ON P_2

In this section, the RATP for (ψ, P_{10}, P_{20}) is addressed relative to the model class \mathcal{D} in which D_2 can vary freely around D_{20} while N_1 and N_2 are held fixed, i.e.,

$$\mathcal{D} \triangleq \{(\dot{N}_{10}D_2^{-1}, N_{20}D_2^{-1}) \in \mathcal{M}(\mathbb{R}_p) \times \mathcal{M}(\mathbb{R}_p) : D_2 \in \mathcal{M}(\mathcal{S}), N_{20} \text{ and } D_2 \text{ RCP}\}.$$

To this effect, note first that $C(R, K)$ solves the RATP for (ψ, P_{10}, P_{20}) relative to \mathcal{D} if and only if R satisfies condition a) and $\dot{N}_{10}\dot{D}_C(K)\Delta D_2(I + \dot{D}_C(K)\Delta D_2)^{-1}\psi^{-1} \in \mathcal{M}(\mathcal{S})$, for all ΔD_2 such that $\|\Delta D_2\|_\infty < \epsilon$ for some $\epsilon > 0$ (this follows from Proposition 1). Due to Lemma 1, the aforementioned condition is equivalent to $\dot{N}_{10}\dot{D}_C(K)\varphi^{-1} \in \mathcal{M}(\mathcal{S})$, i.e., $N_\varphi D_\varphi^{-1}(K\dot{N}_{20} + Y_0) \in \mathcal{M}(\mathcal{S})$, which is equivalent to $D_\varphi^{-1}(K\dot{N}_{20} + Y_0) \in \mathcal{M}(\mathcal{S})$, i.e., there exists $\bar{D}_C \in \mathcal{M}(\mathcal{S})$ such that $-D_\varphi\bar{D}_C + K\dot{N}_{20} = -Y_0$. This equation is formally identical to (1); thus, the arguments which led to Theorem 2 apply *mutatis mutandis* to it and lead to the following theorem.

Theorem 3: Let (ψ, P_{10}, P_{20}) be as above.

- The set C_2 of all controllers which solve the RATP for (ψ, P_{10}, P_{20}) relative to \mathcal{D} is nonempty iff \dot{N}_{10} and ψ are ESP and D_φ and N_{20} are ESP.
- In this case, C_2 is given by

$$C_2 = \{(K\dot{N}_{20} + Y_0)^{-1}[R(K\dot{D}_{20} - X_0)] : (R, K) \in S_2, \det\{(K\dot{N}_{20} + Y_0)(\infty)\} \neq 0\}$$

where $S_2 = \{(R, K) : R = \dot{X}\dot{I} - D, \dot{R}, K = -Y_0\dot{Y}_0 - \dot{D}_\varphi\dot{K}, (\dot{R}, \dot{K}) \in \mathcal{M}(\mathcal{S}) \times \mathcal{M}(\mathcal{S})\}$, the expression for R is the same as in Theorem 2,

$$M_D = \text{GLCD}\{(D_\varphi^{-1} - I), (I - \dot{N}_{20}^I)\}, (D_\varphi^{-1} - I) = M_D \bar{D}_\varphi^{-1}, (I - \dot{N}_{20}^I) = M_D \bar{N}_{20},$$

with $(\bar{D}_\varphi, \bar{N}_{20})$ LCP, \bar{Y}_0, Y_D , and \hat{D}_φ are defined by $Y_0 = M_D \bar{Y}_0$

$$\bar{D}_\varphi X_D + \bar{N}_{20} Y_D = I \quad \text{and} \quad \bar{N}_{20} \hat{D}_\varphi = \bar{D}_\varphi \hat{N}_{20},$$

for some $X_D \in \mathcal{M}(\mathbf{S})$, $\hat{N}_{20} \in \mathcal{M}(\mathbf{S})$. \square

Remark: If \hat{N}_{10} and ψ_r are ESP and $\hat{N}_{10}^{-1} I, I, \psi_r^T$ are RCP, the expression for R in Theorems 2 and 3 does not require Kronecker products (see [14]). Analogously, if $\hat{D}_\varphi = I$ and I, \hat{D}_{20}^T , respectively, I, \hat{N}_{20}^T are RCP, no Kronecker products are needed in the expression for K in Theorem 2, respectively, Theorem 3. \square

Remark: It is worth noting that since $\hat{D}_\varphi = K \hat{N}_{20} + Y_0$ and $\hat{N}_\varphi = K \hat{D}_{20} - X_0$ are LCP, there is no controller which solves the RATP when free, independent perturbations on both N_2 and D_2 are allowed while \hat{N}_1 is held fixed.

VI. PERTURBATIONS ON D_1, M , AND N_1

Perturbations are now considered which are defined on the basis of the coprime factors of P_1 . To this effect, write $P_1 = N_1 D_1^{-1}$ and $P_2 = N_2 D_2^{-1}$ RCP factorizations—due to stabilizability requirements (Theorem 1), $D_2 = D_1 M$, so that $P_1 = \hat{N}_1 D_2^{-1}$, $\hat{N}_1 = N_1 M$.

The RATP for (ψ_r, P_{10}, P_{20}) is now addressed relative to the model class \mathcal{D}_1 , in which D_1 can vary freely around D_{10} while N_1, N_2 , and M are held fixed, i.e.,

$$\mathcal{D}_1 = \{(N_{10} D_1^{-1}, N_{20} (D_1 M_0)^{-1}) \in \mathcal{M}(\mathbf{R}_p) \times \mathcal{M}(\mathbf{R}_p);$$

$$D_1 \in \mathcal{M}(\mathbf{S}), (N_{10}, D_1) \text{RCP}, (N_{20}, D_1 M_0) \text{RCP}\}.$$

Note first that, in this case, $\Delta \hat{N}_1 = 0$, $\Delta N_2 = 0$ and $\Delta D_2 = \Delta D_1 M_0$; thus, it follows from Proposition 1 (and [7, Lemma 5.7.6]), that if $C(R, K)$ solves the RATP for (ψ_r, P_{10}, P_{20}) relative to \mathcal{D}_1 then

$$\begin{aligned} & \hat{N}_{10} \hat{D}_\varphi(K) \Delta D_1 M_0 \\ & \cdot (I + \hat{D}_\varphi(K) \Delta D_1 M_0)^{-1} \psi_r^{-1} \hat{R} \in \mathcal{M}(\mathbf{S}) \\ & \Leftrightarrow \hat{N}_{10} \hat{D}_\varphi(K) \Delta D_1 (I + M_0 \hat{D}_\varphi(K) \Delta D_1)^{-1} \\ & \cdot (M_0 R) \psi_r^{-1} \in \mathcal{M}(\mathbf{S}). \end{aligned}$$

Writing $(M_0 R) \psi_r^{-1} = \hat{\psi}_{r,m}^{-1} M_{r,0}$ (coprime factorizations), it follows from Lemma 1 that the last condition is verified if and only if

$$\hat{N}_{10} \hat{D}_\varphi(K) \varphi_r^{-1} = N_{10} (K \hat{N}_{20} + Y_0) \varphi_r^{-1} \in \mathcal{M}(\mathbf{S})$$

(recall that under the nominal tracking condition $(M_0 R)$ and ψ_r are RCP, so that the greatest invariant factors of $\psi_{r,0}, \psi_{r,m}$, and ψ_r are the same φ_r). This is the same condition found in Section IV, so that Theorem 3 applies here with D replaced by D_1 .

Consider now perturbed model classes \mathcal{D}_{M1} in which M can vary freely around M_0 while N_1, N_2 , and D_1 are held fixed, i.e.,

$$\begin{aligned} \mathcal{D}_{M1} = \{ & (N_{10} D_{10}^{-1}, N_{20} (D_{10} M)^{-1}) \in \mathcal{M}(\mathbf{R}_p) \times \mathcal{M}(\mathbf{R}_p); \\ & M \in \mathcal{M}(\mathbf{S}), (N_{20}, D_{10} M) \text{RCP}\}. \end{aligned}$$

$C(R, K)$ solves the RATP for (ψ_r, P_{10}, P_{20}) relative to \mathcal{D}_{M1} iff R satisfies condition a) of Proposition 1 and

$$\begin{aligned} & (N_{10} \Delta M - N_{10} M_0 \hat{D}_\varphi(K) D_{10} \Delta M) (I + \hat{D}_\varphi(K) D_{10} \Delta M)^{-1} \\ & \cdot \hat{\psi}_{r,m}^{-1} \in \mathcal{M}(\mathbf{S}) \quad \text{for all } \Delta M \end{aligned}$$

such that $\|D_{10} \Delta M\| < \epsilon$, for some $\epsilon > 0$ (this is due to Proposition 1). The last condition is equivalent, by Lemma 1, to $N_{10} (I - M_0 \hat{D}_\varphi(K) D_{10}) \varphi_r^{-1} \in \mathcal{M}(\mathbf{S})$. Let $(\hat{N}_\varphi, \hat{D}_\varphi)$ RCP be defined by $N_{10} \varphi_r^{-1} = \hat{N}_\varphi \hat{D}_\varphi^{-1}$; then the condition above holds iff

$$\begin{aligned} & \text{there exist } K \text{ and } Z \text{ in } \mathcal{M}(\mathbf{S}) \text{ such that } M_0 \hat{D}_\varphi(K) D_{10} \\ & + \hat{D}_\varphi Z = I. \quad (2) \end{aligned}$$

This, in turn, implies that \hat{D}_φ, M_0 LCP and \hat{D}_φ, D_{10} ESP, which by Lemma 2, are equivalent to the existence of $(X, Y) \in \mathcal{M}(\mathbf{S}) \times \mathcal{M}(\mathbf{S})$ such that $M_0 X D_{10} + \hat{D}_\varphi Y = I$. Conversely, suppose that there exists such (X, Y) ; then $M_0 X D_{10} (I - M_0 Y_0 D_{10}) + \hat{D}_\varphi [Y (I - M_0 Y_0 D_{10})] = I - M_0 Y_0 D_{10}$ or, equivalently, $M_0 X D_{10} - M_0 X (D_{10} M_0 Y_0) + \hat{D}_\varphi Z_0 = I - M_0 Y_0 D_{10}$, where $Z_0 \triangleq Y (I - M_0 Y_0 D_{10})$. Now, $D_{10} M_0 Y_0 = D_{20} Y_0 = I - \hat{Y}_0 \hat{N}_{20}$ so that

$$M_0 X D_{10} - M_0 X D_{10} + M_0 X \hat{Y}_0 \hat{N}_{20} D_{10} + \hat{D}_\varphi Z_0 = I - M_0 Y_0 D_{10}$$

or, equivalently,

$$M_0 (X \hat{Y}_0) N_{20} D_{10} + \hat{D}_\varphi Z_0 = I - M_0 Y_0 D_{10},$$

$$M_0 [(X Y_0) N_{20} + Y_0] D_{10} + \hat{D}_\varphi Z_0 = I.$$

It has thus been established that " $\exists (K, Z) \in \mathcal{M}(\mathbf{S}) \times \mathcal{M}(\mathbf{S})$ such that $M_0 \hat{D}_\varphi(K) D_{10} + \hat{D}_\varphi Z = I$ iff $\exists (X, Y) \in \mathcal{M}(\mathbf{S}) \times \mathcal{M}(\mathbf{S})$ such that $M_0 X D_{10} + \hat{D}_\varphi Y = I$ ". By Lemma 2, it then follows that there exists such (K, Z) if and only if M_0 and \hat{D}_φ are LCP and \hat{D}_φ and D_{10} are ESP.

The following theorem can then be stated.

Theorem 4: Let (ψ_r, P_{10}, P_{20}) be as above.

a) The set C_1 of all controllers which solve the RATP for (ψ_r, P_{10}, P_{20}) relative to \mathcal{D}_{M1} , is nonempty iff \hat{D}_φ and M_0 are LCP and \hat{D}_φ and D_{10} are ESP.

b) In this case, C_1 is given by

$$\begin{aligned} C_1 = \{ & (K \hat{N}_{20} + Y_0)^{-1} [R : K \hat{D}_{20} - X_0] : (R, K) \in S_1, \\ & \det \{(K N_{20} + Y_0)(\infty)\} \neq 0 \} \end{aligned}$$

where $S_1 \triangleq \{(R, K) \in \mathcal{M}(\mathbf{S}) \times \mathcal{M}(\mathbf{S}) : R = \hat{N}_1 \hat{I} - D_1 R, K = Y_{M1} X_0 - \hat{D}_{\varphi 1} K \text{ for some } (\hat{R}, \hat{K}) \in \mathcal{M}(\mathbf{S}) \times \mathcal{M}(\mathbf{S})\}$, the expression for R is the same as in Theorem 2,

$$M_{M1} = \text{GLCD} \{(D_\varphi - I), (M_0 - D_{10} \hat{N}_{20}^T)\},$$

$$(D_\varphi - I) = M_{M1} \hat{D}_{\varphi 1}, \quad (M_0 - D_{10} \hat{N}_{20}^T) = M_{M1} N_{20}$$

(which implies \hat{D}_φ, N_{20} LCP), \hat{X}_0, Y_{M1} , and $\hat{D}_{\varphi 1}$ are defined by

$$I - (M_0 - D_{10}^T) Y_0 = M_{M1} X_0, \quad D_{M1} X_{M1} + \hat{N}_{20} Y_{M1} = I,$$

and

$$\hat{N}_{20} \hat{D}_{\varphi 1} = \hat{D}_{\varphi 1} \hat{N}_{20} \quad \text{for some } X_{M1} \in \mathcal{M}(\mathbf{S}), \hat{N}_{20} \in \mathcal{M}(\mathbf{S}).$$

Proof: The proof is exactly like the proof of Theorem 2 when \mathcal{N} and the equation " $K \hat{D}_{20} - D_\varphi \hat{N}_{20} = X_0$ " are replaced, respectively, by \mathcal{D}_{M1} and (2), i.e., " $M_0 K \hat{N}_{20} D_{10} + \hat{D}_\varphi Z = I - M_0 Y_0 D_{10}$ ". \square

Remark: As in the last Remark in Section IV, there is no controller which solves the RATP when free independent perturbations on M and D_1 are allowed while N_1 is held fixed.

Remark It can be shown that if D_φ and M_0 are LCP, then \hat{D}_φ and D_{10} are ESP if and only if D_φ and D_{20} are ESP (the last condition also appears in Theorem 2)

Finally, perturbations on N_1 are considered with D_1, N_2 , and M held fixed. In this case, arbitrary perturbations on N_1 lead to RATPS with no solutions and the most general perturbed model class for which the RATP has a solution is given by

$$\mathcal{V}_\lambda^0(R_0) = \{(N_1 D_{10}^{-1}, N_{20} D_{20}^{-1}) \mid V_1 = V_{10} + S \psi_{10}, S \in \mathcal{M}(\mathcal{S})\}$$

where $\psi_{10} M^{-1} = M_1^{-1} \psi_{10}$, $\psi_{10} R_0 = R_0 \psi_{10}$, for some R_0 satisfying condition a) of Proposition 1 (see [14]), where all controllers that solve the RATP for (ψ_1, P_{10}, P_{20}) relative to $\mathcal{V}_\lambda^0(R_0)$ are explicitly parameterized. An entirely similar situation occurs when perturbations on N_1 are considered with N_2 and D_2 are held fixed (cf. [14])

VII. CONCLUDING REMARKS

In this note, the robust asymptotic tracking problem was addressed for model perturbation classes obtained by allowing one of the coprime factors of the transfer functions from u to y and v to vary while the others are held fixed. In each such a case, explicit, necessary, and sufficient conditions on the problem data are given for the existence of solutions, and, under these conditions, an explicit parameterization of all solutions to each of these problems has been obtained.

APPENDIX

Lemma 1 Let A, B, C, M, Q be matrices in $\mathcal{M}(\mathcal{S})$, ψ_1 non singular biproper, and let φ be the greatest invariant factor of ψ_1 . Then $A(B\Delta\psi + C\Delta D)(I + M\Delta\psi + Q\Delta D)^{-1} \in \mathcal{M}(\mathcal{S})$ for

any $(\Delta\psi, \Delta D)$ such that $\|[\Delta\psi^T, \Delta D^T]^T\|_\infty < \epsilon$ for some $\epsilon > 0$ if and only if $A B \varphi^{-1} \in \mathcal{M}(\mathcal{S})$, $A C \varphi^{-1} \in \mathcal{M}(\mathcal{S})$. If B, C are LCP, the condition above is equivalent to $A \varphi^{-1} \in \mathcal{M}(\mathcal{S})$.

Lemma 2 Let A, B, C be matrices in $\mathcal{M}(\mathcal{S})$. Then there exist X and Y (matrices in $\mathcal{M}(\mathcal{S})$) such that $A \psi B + Y C = I$ if and only if A, C are ESP and B, C are RCP.

Lemma 3 Let A, B, C, D , and Q be matrices in $\mathcal{M}(\mathcal{S})$. The equation $A \psi B + D Y C = Q$ has a solution $(X, Y) \in \mathcal{M}(\mathcal{S}) \times \mathcal{M}(\mathcal{S})$ if and only if $M = \text{GCLD}\{(A, B^T), (D, C^T)\}$ is a left divisor of Q (the column form of Q), i.e., there is a \mathcal{S} -matrix \hat{Q} such that $Q = M \hat{Q}$. In this case, writing $(A, B^T) = M \hat{A}, (D, C^T) = M \hat{C}$ then \hat{A}, \hat{C} are LCP and all pairs (X, Y) in the set defined by $X = \hat{A} \hat{Q} + C K, Y = \hat{C} \hat{Q} - A K, \forall K \in \mathcal{M}(\mathcal{S})$, are solutions of $A \psi B + D Y C = Q$ where X, Y, A and C are \mathcal{S} -matrices such that $A X + C Y = I, A C = C A$ (which exist by the left-coprimeness of A, C). Moreover, if M is nonsingular this is indeed the set of all solutions.

REFERENCES

- [1] E. J. Davison, "The output control of linear time invariant multivariable systems with unmeasurable arbitrary disturbances," *IEEE Trans Automat Contr*, vol. 17, pp. 621–629, 1972.
- [2] —, "The robust control of a servomechanism for a linear time invariant multivariable system," *IEEE Trans Automat Contr*, vol. 21, pp. 25–34, 1976.
- [3] W. M. Wonham, *Linear Multivariable Control*. New York: Springer, 1974.
- [4] B. A. Francis and W. M. Wonham, "The internal model principle for linear multivariable systems," *Appl Math Optimiz*, vol. 2, pp. 170–194, 1975.

- [5] B. A. Francis and Vidyasagar, "Algebraic and topological regulator problem for lumped linear systems," *Automatica*, pp. 87–90, 1983.
- [6] C. A. Desoer and A. N. Gündes, "Algebraic design of linearizable feedback systems," in *Proc IMSE*, 1985, 1985.
- [7] M. Vidyasagar, *Control System Synthesis*. Cambridge, MA: MIT Press, 1985.
- [8] T. Sugie and T. Yoshikawa, "General solution of robust tracking problem in two degree of freedom control systems," *IEEE Trans Automat Contr*, vol. 31, pp. 552–554, 1986.
- [9] S. Hara, "Parametrization of stabilizing controllers for multivariable servo systems with two degrees of freedom," *Int J Contr*, vol. 47, pp. 779–790, 1986.
- [10] S. Hara and T. Sugie, "Independent parameterization of two degree of freedom compensators in general robust tracking systems," *IEEE Trans Automat Contr*, vol. 33, pp. 59–67, 1988.
- [11] C. N. Nett, "Algebraic aspects of linear control system stability," *IEEE Trans Automat Contr*, vol. 34, pp. 1169–1172, 1986.
- [12] P. G. Ferreira, "Four input four output feedback systems: Robust asymptotic behavior," *IEEE Trans Automat Contr*, vol. 33, pp. 1169–1172, 1988.
- [13] T. Sugie and M. Vidyasagar, "Further results on the robust tracking problem in two degree-of-freedom control systems," *Syst Contr Lett*, vol. 13, pp. 101–108, 1989.
- [14] G. O. Correa and M. A. da Silveira, "On robust asymptotic tracking: Perturbations on coprime factors and parameterization of all solutions," Tech. Rep. 41/92, INCC CNPq, Rio de Janeiro, 1992.

A Version of Hautus' Test for Tandem Connection of Linear Systems

Andrea Bacciotti and Giannina Beccari

Abstract—In this note, we consider systems which can be decomposed as a cascade connection of linear subsystems. The second subsystem is assumed to have the controller canonical form. The main result is a version of Hautus' test, which allows us to check controllability and stabilizability properties of these systems by means of reduced-order rank computations.

1. INTRODUCTION

The study of structural properties of interconnected systems is a classical subject (see, for instance, [2], [4], [6], [11]) which has been recently revitalized ([1], [8], [10]). In this note we are interested in autonomous linear control systems of the form

$$\begin{cases} \dot{y} = Ay + B \\ \dot{z} = Cz + Gu \end{cases} \quad (1)$$

where $y \in \mathbb{R}^l$, $z \in \mathbb{R}^q$ denote the state variables and $u \in \mathbb{R}^m$ is the input ($m \leq q$). They can be considered as the tandem (or cascade) connection of the two subsystems

$$\dot{y} = Ay + Bu \quad (1a)$$

(with input $u \in \mathbb{R}^m$) and

$$\dot{z} = Cz + Gu \quad (1b)$$

Manuscript received June 24, 1993; revised December 8, 1993 and March 18, 1994.

The authors are with the Dipartimento di Matematica del Politecnico, 10129 Torino, Italy.

IEEE Log Number 9407020

the connection being established setting $v = z$.

In the sequel, we denote by $\mathbb{R}^{\nu, \mu}$ (respectively, $\mathbb{C}^{\nu, \mu}$) the space of real (respectively, complex) matrices with ν rows and μ columns. Accordingly, $A \in \mathbb{R}^{p, p}$, $B \in \mathbb{R}^{p, q}$, $F \in \mathbb{R}^{q, q}$, and $G \in \mathbb{R}^{q, m}$.

It is well known that if (1) is controllable [stabilizable] then (1a) and (1b) are separately controllable [stabilizable] (see [4]). However, the converse is false in general (see again [4] and the simple example in Section II). Thus, it is natural to ask what kind of condition should be imposed on (1a) and (1b) in order to guarantee controllability [stabilizability] of the overall system (1).

It is easy to see that if $q = m$ and G is not singular, then (1) is controllable [stabilizable] if and only if (1a) is controllable [stabilizable]. However, this assumption is too conservative to be interesting. From a more general point of view, the problem is addressed in [6], [2] and especially in [4] using a state-space description similar to (1), and in [11], [3] using differential operator representation and coprime factorization. We note that the necessary and sufficient conditions obtained in [4] require some restrictions about the Jordan canonical form of matrices A and F . For instance, it is assumed that there is only one Jordan block for each possible common eigenvalue of A and F .

In this note, we assume that

$$F = \text{diag}(C_1, \dots, C_m), \quad G = \text{diag}(g_1, \dots, g_m) \quad (2)$$

where for each $i = 1, \dots, m$

$$C_i = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{k_i, k_i}$$

and

$$g_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \in \mathbb{R}^{k_i, 1} \quad (3)$$

($k_1 + \dots + k_m = q$). In other words, we assume that (1b) is in Brunowski's canonical form (see [9, p. 139]). From a practical point of view, we can think of (1), (2), and (3) as a plant represented by the linear system (1a) and coupled with m parallel chains of integrators of (possibly) different lengths.

Chains of integrators are easy to implement. A typical situation occurs when (1a) is given and one has to design a dynamical compensator. Usually, in these cases only the first component of the state vector z affects directly the first subsystem. Moreover, the form (1), (2), (3) incorporates systems whose behavior depends on the derivatives of the input map. These possible applications will be shortly discussed in Section IV.

When (1b) is in Brunowski's canonical form, then F has zero as a unique eigenvalue. Since $m \geq 1$ and zero is allowed to be an eigenvalue of A as well, our result applies also to cases not covered by [4]. Finally, we note that if (1b) is controllable, then the structure (2) and (3) can be recovered up to a linear change of coordinates

$$\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} I_p & O \\ O & T \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} \quad (4)$$

(where $T \in \mathbb{R}^{q, q}$), a preliminary linear feedback transformation and the elimination of possible redundant input channels ([9]). Thus, the assumption that (1b) is in Brunowski's canonical form turns out to be restrictive only when we are interested in the stabilizability problem.

Of course, we can think of system (1), (2), (3) as a whole system. To this purpose, we set $n = p + q$, $x = (y, z)$

$$\dot{A} = \begin{pmatrix} A & B \\ O & F \end{pmatrix}, \quad \dot{B} = \begin{pmatrix} O \\ G \end{pmatrix} \quad (5)$$

where we denote by O some matrices with zero entries. Accordingly, its properties can be investigated with the aid of some classical criterion. For instance, Hautus' test (see [9, p. 93, 153]) states that the system is controllable if and only if

$$\text{rank}[\dot{A} - \lambda I_n, \dot{B}] = n \quad (6)$$

for each $\lambda \in \mathbb{C}$, and stabilizable if and only if (6) holds for each $\lambda \in \mathbb{C}_0^+ = \{z \in \mathbb{C} : \text{Re } z \geq 0\}$. Note that the matrix in (6) has n rows and $n + m$ columns.

However, as remarked in [4], with this approach, it is difficult to state the conditions in terms of the subsystems. Moreover, disregarding the structure, there is no gain in the process insight.

The main contribution of this note is a version of Hautus' test which, taking into account the form of the system, allows us to check the controllability and stabilizability properties by computing the rank of a reduced-order matrix, namely a matrix with p rows and $p + m$ columns. The result is stated in Section II and proved in Section III. Additional comments and remarks are finally provided in Section IV.

II. THE RESULT

As already mentioned, if (1a) and (1b) are both controllable [stabilizable], (1) is not necessarily controllable [stabilizable].

Example: Consider the following system with $p = 1$, $q = 2$:

$$\begin{cases} \dot{y} = y + z_1 - z_2 \\ \dot{z}_1 = z_2 \\ \dot{z}_2 = u. \end{cases}$$

Both subsystems are controllable, but the overall system is not even stabilizable.

In the following theorem a special role is played by the columns of B . We introduce some notation. Let us write $B = (B_1, \dots, B_m)$, where for each $i = 1, \dots, m$

$$B_i = [b_{i,1}, b_{i,2}, \dots, b_{i,k_i}] \in \mathbb{R}^{p, k_i}.$$

Moreover, let

$$h_i(\lambda) = b_{i,1} + \lambda b_{i,2} + \lambda^2 b_{i,3} + \dots + \lambda^{k_i-1} b_{i,k_i}$$

and

$$H(\lambda) = [h_1(\lambda), \dots, h_m(\lambda)] \in \mathbb{C}^{p, m}.$$

Now, the main result of this note can be stated.

Theorem: Let \dot{A}, \dot{B} be given by (5), (2), (3), and let $\lambda \in \mathbb{C}$. Then, $\text{rank}[\dot{A} - \lambda I_n, \dot{B}] = n = p + q$, i.e., maximal, if and only if

$$\text{rank}[A - \lambda I_p, H(\lambda)] = p. \quad (7)$$

i.e., maximal.

It follows that System (1), (2), (3) is controllable [stabilizable] if and only if (7) holds for each $\lambda \in \mathbb{C}$ [for each $\lambda \in \mathbb{C}_0^+$]. As explained in the Introduction, we have so obtained a criterion which allows us to investigate the controllability and stabilizability properties of the overall system by means of reduced-order matrix rank computations. This is illustrated by the following example.

Example: Let us consider a system of the form (1), (2), (3) with

$$A = \begin{pmatrix} 3 & 3 \\ 4 & 7 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & k & 1 \\ 2 & 4 & 0 \end{pmatrix}$$

where k is a real parameter. Let $m = 1$, so that

$$\tilde{A} = \begin{pmatrix} 3 & 3 & 1 & k & 1 \\ 4 & 7 & 2 & 4 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \tilde{B} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

In order to apply Hautus' test to the pair (\tilde{A}, \tilde{B}) , we should compute the rank of $(\tilde{A} - \lambda I_5, \tilde{B})$ for each eigenvalue λ of \tilde{A} (actually, it is not difficult to see that we can limit ourselves to the eigenvalues of A). The eigenvalues of A are, in this case, $\lambda = 1$ and $\lambda = 9$. Let us take for instance $\lambda = 1$. The matrix to be considered is

$$\begin{pmatrix} 2 & 3 & 1 & k & 1 & 0 \\ 4 & 6 & 2 & 4 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}.$$

Clearly, the rank of this matrix is maximal if and only if the 4×4 matrix obtained by deleting the first column, the last column, and the last row is nonsingular. The problem is so reduced to the (borrowing) computation of the determinant of a fourth-order matrix.

Using our criterion, one is led to consider the simpler matrix

$$(A - I, b_0 + b_1 + b_2) = \begin{pmatrix} 2 & 3 & 2+k \\ 4 & 6 & 6 \end{pmatrix}.$$

It is immediate to see that its rank is maximal if and only if $k \neq 1$.

III. THE PROOF

Let $M = [\tilde{A} - \lambda I_n, \tilde{B}]$. In order to compute its rank, it is convenient to multiply M on the right by an invertible matrix Q . The construction of Q will be accomplished in three steps.

In what follows, we denote by $O_{\nu, \mu}$ the matrix with ν rows, μ columns, and zero entries (to simplify the notation, the subscripts ν, μ will be suppressed whenever the matrix dimensions are clear from the context).

Step 1: It is convenient to look at M as a block-matrix of the form

$$[M_0, M_1, \dots, M_m, N_1, \dots, N_m],$$

where

$$M_0 = \begin{pmatrix} A - \lambda I_p \\ O_{q, p} \end{pmatrix} \in \mathbb{C}^{n, p},$$

$$M_i = \begin{pmatrix} B_i \\ O_{k_i, k_i} \\ \vdots \\ C_i - \lambda I_{k_i} \\ \vdots \\ O_{k_m, k_i} \end{pmatrix} \in \mathbb{C}^{n, k_i} \quad \text{and} \quad N_i = \begin{pmatrix} O_{p, 1} \\ O_{k_1, 1} \\ \vdots \\ g_i \\ \vdots \\ O_{k_m, 1} \end{pmatrix} \in \mathbb{C}^{n, 1}$$

($i = 1, \dots, m$). We perform a reordering of the blocks of M , by multiplying on the right by a suitable invertible matrix Q^1 . It is convenient to define Q^1 as a block matrix. More precisely, we set $Q^1 = (Q^1_{ij})(i, j = 0, 1, \dots, 2m)$ where

$$Q^1_{0,0} = I_p,$$

$$Q^1_{i, 2i-1} = I_{k_i} \quad \text{and} \quad Q^1_{s+m, 2s} = 1 \quad \text{for } r, s = 1, \dots, m$$

$$Q^1_{i,j} = 0 \quad \text{otherwise.}$$

As a result, we obtain $MQ^1 = [M_0, M_1, N_1, \dots, M_m, N_m]$.

Step 2: Let us multiply MQ^1 on the right by another invertible matrix Q^2 . In order to describe Q^2 , it is convenient to think on Q^2 as a block-matrix in a different way

$$MQ^1 = \begin{pmatrix} A' & B' \\ O & F' \end{pmatrix}$$

where

$$A' = A - \lambda I_p,$$

$$B' = [B'_1, \dots, B'_m] \quad \text{and} \quad B'_i = [B_i, O_{p, 1}] \in \mathbb{C}^{p, k_i+1},$$

$$F' = \text{diag}(C'_1, \dots, C'_m)$$

and

$$C'_i = [C_i - \lambda I_{k_i}, g_i] = \begin{pmatrix} -\lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & -\lambda & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -\lambda & 1 \end{pmatrix} \in \mathbb{C}^{k_i, k_i+1}.$$

The matrix Q^2 has the form

$$Q^2 = \begin{pmatrix} I_p & O \\ O & \Lambda \end{pmatrix}$$

where $\Lambda = \text{diag}(\Lambda_1, \dots, \Lambda_m) \in \mathbb{C}^{q+m, q+m}$ and

$$\Lambda_i = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \lambda & 1 & 0 & \dots & 0 \\ \lambda^2 & \lambda & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \lambda^{k_i} & \lambda^{k_i-1} & \lambda^{k_i-2} & \dots & 1 \end{pmatrix} \in \mathbb{C}^{k_i+1, k_i+1}.$$

As a consequence

$$MQ^1 Q^2 = \begin{pmatrix} A' & B' \Lambda \\ O & F' \Lambda \end{pmatrix} \quad (8)$$

where

$$B' \Lambda = [B'_1 \Lambda_1, \dots, B'_m \Lambda_m] = [B''_1, \dots, B''_m],$$

$$B''_i = (b''_{i,1}, b''_{i,2}, \dots, b''_{i,k_i}, O_{p,1}) \in \mathbb{C}^{p, k_i+1},$$

$$b''_{i,r} = b_{i,r} + \lambda b_{i,r+1} + \lambda^2 b_{i,r+2} + \dots + \lambda^{k_i-r} b_{i,k_i}$$

and

$$F' \Lambda = \text{diag}(J_1, \dots, J_m), \quad J_m = [O_{k_m, 1}, I_{k_m}] \in \mathbb{C}^{k_m, k_m+1}.$$

The result of this second transformation is that the dependence on λ has been eliminated in the second row of (8).

Step 3: Finally, let us rewrite $MQ^1 Q^2$ as a block matrix with one row

$$MQ^1 Q^2 = [M_0, H_1, K_1, \dots, H_m, K_m]$$

with

$$H_i = \begin{pmatrix} h_i(\lambda) \\ O_{q, 1} \end{pmatrix},$$

$$h_i(\lambda) = b_{i,1} + \lambda b_{i,2} + \lambda^2 b_{i,3} + \dots + \lambda^{k_i-1} b_{i,k_i}$$

and

$$K_i = \begin{pmatrix} b''_{i,2} & \dots & b''_{i,k_i} & O_{p,1} \end{pmatrix}.$$

The purpose of the third transformation is to provide a further reordering of the columns of MQ^1Q^2 . Setting $Q^3 = (Q_{i,j}^3)(i, j = 0, 1, \dots, 2m)$ where

$$Q_{0,0}^3 = I_p$$

$$Q_{2r-1,r}^3 = 1, \quad \text{and} \quad Q_{2s,s+m}^3 = I_{h_s} \quad \text{for } r, s = 1, \dots, m,$$

$$Q_{i,j}^3 = 0 \quad \text{otherwise,}$$

we obtain $MQ^1Q^2Q^3 = [M_0, H_1, \dots, H_m, K_1, \dots, K_m]$.

We are now ready to get the conclusion. Let $Q = Q^1Q^2Q^3$. It is clear that

$$[\tilde{A} - \lambda I_n, \tilde{B}]Q = \begin{pmatrix} A - \lambda I_p & H(\lambda) & L(\lambda) \\ O_{q,p} & O_{q,m} & I_q \end{pmatrix}$$

where $L(\lambda)$ is some matrix which we need not specify. As an immediate consequence, we see that

$$\text{rank}[\tilde{A} - \lambda I_n, \tilde{B}] = \text{rank}[\tilde{A} - \lambda I_n, \tilde{B}]Q = n$$

if and only if $\text{rank}[A - \lambda I_p, H(\lambda)] = p$.

We note that in the case $m = 1$, we have $Q^1 = Q^3 = I_n$. In other words, if $m = 1$ no reordering is needed.

IV. FINAL REMARKS

Condition (7) implies, in particular, that

$$\text{rank}[A - \lambda I_p, B] = p$$

and so we recover the well-known fact that controllability [stabilizability] of (1a) is necessary for controllability [stabilizability] of (1). More precisely, let us define the deficiency of A relative to a set $S \subseteq \mathbb{C}$ as

$$d(A, S) = p - \min_{\lambda \in S} \text{rank}[A - \lambda I_p].$$

According to our theorem, a necessary condition for controllability [stabilizability] of (1) is that

$$d(A, \mathbb{C}) \leq m, \quad \text{respectively } d(A, \mathbb{C}_0^+) \leq m.$$

This conclusion is not surprising. Indeed, if we want to control (1a) and the control inputs are implemented through (1b), then it should be actually possible to control (1a) by means of $m \leq q$ independent inputs.

Sometimes, systems of the form

$$\begin{cases} \dot{y} = Ay + bz_1 \\ \dot{z}_1 = z_2 \\ \vdots \\ \dot{z}_q = u \end{cases} \quad (9)$$

where $b \in \mathbb{R}^p$ and $z_1, \dots, z_q, u \in \mathbb{R}$, have been considered in the literature ([8]). They represent a situation where the second subsystem is a chain of integrators with output z_1 , and the first subsystem depends only on z_1 . System (9) is a particular case of System (1). (2), (3), with $m = 1$. Hence, it follows from our theorem that (9) is controllable [stabilizable] if and only if the pair (A, b) is controllable [stabilizable].

As already recalled in the Introduction, there are other ways to study system (1). For instance, in [4] a change of coordinates

$$\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} P & O \\ O & T \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix}$$

(more general than (4)) is used in order to put A and F in Jordan canonical form. From a numerical point of view, the difficulties encountered in obtaining Brunowski's canonical form are comparable to those encountered in obtaining Jordan form. Thus, the approach followed in the present note should be considered as an alternative to the approach of [4].

Sometimes, systems involving an input map and its derivatives are considered in the control theory literature (see, for instance, [5], [7]). These systems have the form

$$\dot{y} = Ay + B_0v + B_1\dot{v} + \dots + B_lv^{(l)} \quad (10)$$

where $v \in \mathbb{R}^m$ and $B_i \in \mathbb{C}^{n \times m}$ ($i = 0, \dots, l$). Setting

$$z_1 = v, z_2 = \dot{v}, \dots, z_{l+1} = v^{(l)}, \quad \text{and} \quad v^{(l+1)} = u$$

(10) can be equivalently rewritten as

$$\begin{cases} \dot{y} = Ay + B_0z_1 + \dots + B_lz_{l+1} \\ \dot{z}_1 = z_2 \\ \vdots \\ \dot{z}_{l+1} = u \end{cases}$$

(note that here $z_{i+1} \in \mathbb{R}^m$ for each $i = 0, \dots, l$). Up to a reordering of the variables, (10) appears to be a particular case of (1), (2), (3) and its controllability properties can be tested by means of the Theorem of Section II. It turns out that (10) is controllable [stabilizable] if and only if

$$\text{rank}[A - \lambda I_p, B_0 + \lambda B_1 + \dots + \lambda^l B_l] = p$$

for each $\lambda \in \mathbb{C}$ [for each $\lambda \in \mathbb{C}_0^+$].

A different criterion for controllability of (10) can be found in [5]. It is actually an extension of well-known Kalman's controllability matrix rank condition.

REFERENCES

- [1] A. Bacciotti, P. Boieri, and L. Mazzi, "Linear stabilization of nonlinear cascade systems," *Math. Contr., Signals, Syst.*, vol. 6, pp. 146-165, 1993.
- [2] M. V. Bhandarkar and M. M. Fahmy, "Controllability of tandem connected systems," *IEEE Trans. Automat. Contr.*, vol. 17, pp. 150-151, 1972.
- [3] F. M. Callier and C. D. Nahum, "Necessary and sufficient condition for the complete controllability and observability of systems in series using the coprime factorization of a rational matrix," *IEEE Trans. Circ. Syst.*, vol. 22, pp. 90-95, 1975.
- [4] C. T. Chen and C. A. Desoer, "Controllability and observability of composite systems," *IEEE Trans. Automat. Contr.*, vol. 12, pp. 402-409, 1967.
- [5] L. Dai, "Control problem for linear systems with input derivatives control," *Int. J. Syst. Sci.*, vol. 19, pp. 1645-1653, 1988.
- [6] E. G. Gilbert, "Controllability and observability in multivariable control systems," *SIAM J. Contr.*, vol. 2, pp. 128-151, 1963.
- [7] L. Pandolfi, "Generalized control systems, boundary control systems, and delayed control systems," *Math. Contr., Signals, Syst.*, vol. 3, pp. 165-181, 1990.
- [8] A. Saberi, P. V. Kokotovic and H. J. Sussmann, "Global stabilization of partially linear composite systems," *SIAM J. Contr. Optimiz.*, vol. 28, pp. 1491-1503, 1990.
- [9] E. D. Sontag, *Mathematical Control Theory*. New York: Springer Verlag, 1990.
- [10] H. J. Sussmann and P. V. Kokotovic, "The peaking phenomenon and the global stabilization of nonlinear systems," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 424-440, 1991.
- [11] W. A. Wolovich and H. L. Hwang, "Composite system controllability and observability," *Automatica*, vol. 10, pp. 209-212, 1974.

Poisson LQR Design for Asynchronous Multirate Controllers

Robert P. Leland

Abstract—We derive the optimal LQR controller for asynchronous multirate digital control. Independent processors operating without synchronization update different components of the control vector at Poisson arrival times. An iterative algorithm to compute the steady-state control law and optimal cost is described and its convergence is demonstrated.

I. INTRODUCTION

For complex systems, such as aircraft, it can be desirable to distribute the control task among several processors, rather than using a single central computer. If these processors are not triggered by a common clock pulse, and their computation, sampling, and hold activities are not synchronized, we call them asynchronous controllers. In addition, these processors need not operate with the same sampling rate. Multirate sampling in control systems has been of interest since the 1950s. The sampling rates of the controllers are typically assumed to be integrally proportional, and sampling is synchronized to make the sampling process periodic, with a period equal to an integral multiple of the largest sampling period ([1], [2], [5], [7] and others). We will allow arbitrary sampling rates, and random sampling times for each processor, so sampling is no longer periodic.

We wish to control the continuous-time system

$$\dot{x}(t) = Ax(t) + Bu(t) \quad (1)$$

to minimize the cost functional

$$J = E \int_0^T [Qx(t), x(t)] + [Ru(t), u(t)] dt, \quad Q \geq 0, R > 0. \quad (2)$$

We partition the control vector $u(t) \in R^m$ into $n \leq m$ (possibly vector) components:

$$\begin{aligned} u_1(t) \\ \vdots \\ u_n(t) \end{aligned} \quad (3)$$

and assign an independent processor to each of the components $u_i(t)$. Processor i updates $u_i(t)$ at discrete times, with $u_i(t)$ constant between updates. The processors operate asynchronously, hence the update times of the processors are not coordinated in any way. We also assume the update times of processor i occur randomly as the arrival times of a Poisson process $N_i(t)$, where $N_1(t) \cdots N_n(t)$ are independent Poisson processes with arrival rates $\lambda_1 \cdots \lambda_n$. The output of processor i is described by the stochastic differential equation

$$du_i(t) = dN_i(t)(v_i(t) - u_i(t)). \quad (4)$$

If τ is an arrival time of $N_i(t)$, $v_i(\tau)$ is the control update computed by processor i .

We consider such asynchronous LQR controllers, assuming each processor has access to the entire state and all control signals at the time the control signal is calculated.

Manuscript received February 22, 1993; revised January 12, 1994.

The author is with the Department of Electrical Engineering, University of Alabama, Tuscaloosa, AL 35487 USA.

IEEE Log Number 9407021.

The Poisson process approach, although allowing additional uncertainty in sampling times, removes the need to synchronize sampling and permits multirate sampling with sampling rates that are not integral multiples.

Poisson processes have appeared in control problems in several contexts. A large body of literature exists on the control of jump parameter linear systems, in which the system parameters change at random times, e.g., [6], [12], and [15]. The LQR problem for such systems was solved in [8], [13], and [17], and the LQG controller is derived in [6].

Stability and control of systems effected by Poisson noise are treated in [9] and [17]. Malliavin calculus for systems driven by Poisson processes is described in [3].

In the asynchronous control problem, jumps occur in the state, rather than in the parameters. The LQR controller for systems with jumps in the state due to compound Poisson noise was derived in [17]. An LQR controller for systems with jumps in the state due to switching between linearized system models was derived in [14]. Existence of the optimal control and continuity of the optimal cost for diffusion processes with jumps were shown in [11]. In each case the Poisson processes represented noise, rather than a control update, and the control signal could be updated in continuous time.

Optimal impulsive control, where control decisions are only made at discrete Poisson arrival times, is treated in [4] for a class of problems in operations research, and existence of the optimal control is demonstrated using likelihood ratios. In [10], the existence of an optimal impulsive control for Markov processes is shown, and the cost-to-go, or value function, is shown to be a function of the current state and control values.

In this note, we formulate and solve the LQR problem for asynchronous multirate controllers using stochastic differential equations. We assume the control signal, or part of it, can only be updated at Poisson arrival times, as in the impulse control problem, treated in the general case in [4] and [10]. We consider the specific case of the LQR problem, and obtain explicit expressions for the optimal control and cost. Our problem also differs slightly from that of [10], in that the control task is distributed among several asynchronous processors, and only part of the control vector can be updated at a single arrival time. When a steady-state controller exists, we describe an iterative algorithm to compute the control gain, and demonstrate its convergence. We compute the control gain for a system of three masses with springs and dampers, and compare the cost to that for a continuous time LQR controller.

II. SYSTEM MODEL

The controller described by (4) can be rewritten in vector form as

$$du(t) = dN(t)(v(t) - u(t)) \quad (5)$$

where $v(t) = [v_1(t)^* \cdots v_n(t)^*]^*$ is a vector of potential control signal updates

$$N(t) = \sum_{i=1}^n N_i(t)I_i.$$

I_i is the $m \times m$ projection matrix that projects the control input space R^m onto the subspace of R^m updated by processor i . Thus, $I_i u(t) = [0 \cdots 0 u_i(t)^* 0 \cdots 0]^*$, where $u_i(t)$ is the (possibly vector)

part of $u(t)$ updated by processor i . As an example, if $m = 7$ and $n = 3$,

$$u(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{bmatrix}, \quad I_1 = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

$$I_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad I_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix}.$$

The stochastic differential equation (5) is understood as the limit of

$$u(t+h) - u(t) = (N(t+h) - N(t))(v(t) - u(t))$$

as $h \rightarrow 0$. Thus, each arrival of $N_i(t)$ causes an update of component $u_i(t)$, where $u_i(t)$ is set to $v_i(t)$, newly calculated by processor i , or $u(t)$ is replaced by $u(t) + I_i(v(t) - u(t))$.

We define the Poisson rate matrix Λ as

$$\Lambda = \sum_{i=1}^n \lambda_i I_i. \quad (6)$$

We define an augmented state vector for this system by adding "hold" states

$$X(t) = \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}. \quad (7)$$

Then the entire system can be described by the following stochastic differential equation:

$$dX(t) = \begin{bmatrix} A dt & B dt \\ 0 & -dN(t) \end{bmatrix} X(t) + \begin{bmatrix} 0 \\ dN(t) \end{bmatrix} v(t). \quad (8)$$

Since $X(t)$ is a Markov process, the control updates $v_i(t)$, $i = 1 \dots n$, will only depend on $X(t)$, and not on the time since the last update.

We define a matrix function $\delta(\cdot)$ on $m \times m$ square matrices as

$$\delta(P) = \sum_{i=1}^n I_i P I_i. \quad (9)$$

The function δ keeps only the block diagonal terms of matrix P . $\delta(P)$ can be computed easily using an element-by-element multiplication by a matrix of ones and zeros. A matrix P is suitably block diagonal if $\delta(P) = P$. The following properties hold for such suitably block diagonal matrices: i) $I_i P = P I_i$, ii) $P^{-1} = \delta(P^{-1})$ if P^{-1} exists, iii) $\Lambda P = P \Lambda$, iv) If $P > 0$, then $\delta(P) > 0$.

III. OPTIMAL POISSON LQR CONTROLLER

We derive an optimal Poisson LQR controller using dynamic programming. We assume up-to-date analog measurements of the entire state vector and all control signals are available to each controller at that controller's update times. In the augmented state-space formulation, the cost function (2) may be rewritten as

$$J = E \int_0^T [QX(t), X(t)] dt$$

$$Q = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}.$$

To apply dynamic programming, let the cost-to-go be

$$V(X, t) = E \left[\int_t^T [QX(\tau), X(\tau)] d\tau \mid X(t) = X \right].$$

A necessary condition for optimality is given by the following Hamilton-Jacobi-Bellman equation:

$$0 = \frac{\partial}{\partial t} V(X, t) + \min_v \{ \mathcal{L}_v V(X, t) + [QX, X] \}$$

where \mathcal{L}_v is the generator for the stochastic differential equation (8):

$$\mathcal{L}_v G(X) = \nabla G(X) \cdot AX + \sum_{i=1}^n \lambda_i \left\{ G \left(X - \begin{bmatrix} 0 & 0 \\ 0 & I_i \end{bmatrix} X + \begin{bmatrix} 0 \\ I_i \end{bmatrix} v \right) - G(X) \right\}$$

where we denote

$$A = \begin{bmatrix} A & B \\ 0 & 0 \end{bmatrix}.$$

We assume a solution of the form

$$V(X, t) = [M(T-t)X, X].$$

Let

$$M = \begin{bmatrix} M_1 & M_2 \\ M_2^* & M_3 \end{bmatrix}.$$

Then $\mathcal{L}_v[MX, X]$ can be expressed as

$$\begin{aligned} \mathcal{L}_v[MX, X] &= [(MA + A^*M)X, X] \\ &\quad + \sum_{i=1}^n \lambda_i \left\{ \left[M \left(X - \begin{bmatrix} 0 & 0 \\ 0 & I_i \end{bmatrix} X + \begin{bmatrix} 0 \\ I_i \end{bmatrix} v \right) \right. \right. \\ &\quad \left. \left. \left(X - \begin{bmatrix} 0 & 0 \\ 0 & I_i \end{bmatrix} X + \begin{bmatrix} 0 \\ I_i \end{bmatrix} v \right) \right] - [MX, X] \right\} \\ &= \left\{ \left[(MA + A^*M) - M \begin{bmatrix} 0 & 0 \\ 0 & \Lambda \end{bmatrix} \right. \right. \\ &\quad \left. \left. - \begin{bmatrix} 0 & 0 \\ 0 & \Lambda \end{bmatrix} M + \begin{bmatrix} 0 & 0 \\ 0 & \Lambda \delta(M_3) \end{bmatrix} \right] X, X \right\} \\ &\quad + [\Lambda [M_2^* (M_3 - \delta(M_3))] X, v] \\ &\quad + [v, \Lambda [M_2^* (M_3 - \delta(M_3))] X] \\ &\quad + [\Lambda \delta(M_3) v, v]. \end{aligned}$$

This is minimized by a choice of

$$v = -\delta(M_3)^{-1} [M_2^* (M_3 - \delta(M_3))] X$$

which is the optimal linear feedback control law. Then

$$\begin{aligned} \min_v \mathcal{L}_v[MX, X] &= [(MA + A^*M)X, X] \\ &\quad - \left[M \begin{bmatrix} 0 \\ I \end{bmatrix} \Lambda \delta(M_3)^{-1} [0 \ I] MX, X \right]. \end{aligned}$$

Thus, the Hamilton-Jacobi-Bellman equation is satisfied if $M(T-t)$ satisfies

$$\begin{aligned} 0 &= \frac{\partial}{\partial t} M(T-t) + A^* M(T-t) + M(T-t)A + Q \\ &\quad - M(T-t) \mathcal{E} \Lambda \delta(M_3(T-t))^{-1} \mathcal{E}^* M(T-t) \end{aligned}$$

with endpoint condition $M(0) = 0$, where

$$\mathcal{E} = \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

The optimal control is

$$v(t) = -\delta(M_3(T-t))^{-1} [M_2^*(T-t)^* (M_3(T-t) - \delta(M_3(T-t)))] X(t) \quad (10)$$

and the optimal cost is

$$V(X(0), 0) = [M(T)X(0), X(0)]. \quad (11)$$

Note that $v(t)$ depends on both $x(t)$ and $u(t)$, however $v_i(t)$, the potential update for controller i depends only on the state $x(t)$ and the other control signals $u_k(t)$, $k \neq i$.

IV. STEADY-STATE CONTROLLER

It is difficult to determine the existence of a steady-state controller because the system contains a random matrix multiplied by the state. However, given a solution to the steady-state Riccati equation, we can show the existence of a steady-state control.

Theorem 4.1: Suppose that the steady-state Riccati equation (SSRE)

$$0 = A^*M + MA + Q - M\bar{B}\bar{A}\delta(M_3)^{-1}\bar{B}^*M \quad (12)$$

has a nonnegative solution M' . Then as $T \rightarrow \infty$, $M(T)$ converges monotonically upward to a limit M'' , where M'' satisfies the SSRE (12), and $M'' \leq M'$.

Proof: $M(T)$ is monotonically nondecreasing in T , since $[M(T)X(0), X(0)]$ is the optimal cost for the T time-horizon problem. The minimal cost for the cost functional

$$J' = E \left[\int_0^T [QX(t), X(t)] dt + [M'X(T), X(T)] \right]$$

is $J' = [M'X(0), X(0)]$, since M' is a constant solution to the Riccati equation with initial condition $M(0) = M'$. Since $J' \geq J$, we must have $M' \geq M(T)$ for all T . Since $M(T)$ is bounded above and nondecreasing, $M(T)$ converges to a matrix M'' , and $M'' \leq M'$. \square

By iteratively solving a sequence of steady-state Riccati equations, we can calculate matrix M'' whenever it exists.

Theorem 4.2: Suppose that (A, B) is controllable, (Q, A) is observable, and $R > 0$ is suitably block diagonal $R = \delta(R)$. Suppose also that there exists a nonnegative solution M' to the SSRE (12). Then the steady-state solution to the Riccati equation $M'' = \lim_{T \rightarrow \infty} M(T)$ can be calculated by iteratively solving the Riccati equation

$$0 = A^*M^{(n+1)} + M^{(n+1)}A + Q - M^{(n+1)}\bar{B}\bar{A}\delta(M_3^{(n)})^{-1}\bar{B}^*M^{(n+1)} \quad (13)$$

with $\delta(M_1^{(0)}) = \Lambda^{-1}R$, and the sequence $M^{(n)}$ converges to M'' as $n \rightarrow \infty$.

Proof: Denote by $M = \Phi(P)$ the solution to the Riccati equation

$$0 = A^*M + MA + Q - M\bar{B}\bar{A}P^{-1}\bar{B}^*M \quad (14)$$

for any positive definite suitably block diagonal $m \times m$ matrix $P = \delta(P) > 0$. Also, for the solution M , denote $\delta(M_3) = \Phi_3(P)$. For every $P > 0$, (14) has a unique positive definite solution $M = \Phi(P)$, since (A, \bar{B}) is controllable (implied by (A, B) controllable), and (Q, A) is observable (implied by (Q, A) observable and $R > 0$). Our iterative algorithm is $M^{(n+1)} = \Phi(\delta(M_1^{(n)}))$.

If $P \geq P'$, then by [16, Theorem 1], $\Phi(P) \geq \Phi(P')$, and hence $\Phi_3(P) \geq \Phi_3(P')$. Thus, Φ and Φ_3 are both nondecreasing functions.

We wish to show the sequence $M^{(n)}$ is nondecreasing if we take as our initial condition the following:

$$M^{(0)} = \begin{bmatrix} 0 & 0 \\ 0 & \Lambda^{-1}R \end{bmatrix}$$

Then $M^{(1)} = \Phi(\Lambda^{-1}R)$, and $\Delta M = M^{(1)} - M^{(0)}$, then

$$0 = (A - \bar{B}\bar{A}\bar{B}^*)^* \Delta M + \Delta M(A - \bar{B}\bar{A}\bar{B}^*) - \Delta M\bar{B}\bar{A}R^{-1}\bar{A}\bar{B}^*\Delta M$$

$$Q_0 = \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix}$$

Equation (15) has unique nonnegative definite solution ΔM since $(A - \bar{B}\bar{A}\bar{B}^*, \bar{B})$ is controllable (implied by (A, B) controllable) and $(Q_0, A - \bar{B}\bar{A}\bar{B}^*)$ is detectable (implied by (Q, A) observable). Hence $M^{(1)} \geq M^{(0)}$ and $\delta(M_3^{(1)}) \geq \delta(M_3^{(0)})$.

Performing the next iteration yields: $M^{(2)} = \Phi(\delta(M_1^{(1)})) \geq \Phi(\delta(M_1^{(0)})) = M^{(1)}$, and $\delta(M_3^{(2)}) \geq \delta(M_3^{(1)})$. By induction both $M^{(n)}$ and $\delta(M_3^{(n)})$ are nondecreasing sequences.

To show convergence, we must show that $M^{(n)}$ is uniformly bounded above. By Theorem 4.1, $M(T) \rightarrow M''$, where $M'' \leq M'$ is a solution to the SSRE. Our desired solution can then be written as $M'' = \Phi(\delta(M_1''))$.

We claim that $M_1'' \geq \Lambda^{-1}R$. If $x(0) = 0$, the optimal cost for the infinite time-horizon problem is $[M_1''u(0), u(0)]$. The cost accrued prior to the first update of each processor is

$$[\Lambda^{-1}Ru(0), u(0)] = \sum_{i=1}^n \lambda_i^{-1} [Ri, u(0), I, u(0)]$$

where λ_i^{-1} is the expected value of the first update time of processor i . Hence $M_1'' \geq \Lambda^{-1}R$ since M_1'' describes the cost for a longer time interval. Hence also, $\delta(M_3^{(0)}) = \Lambda^{-1}R \leq \delta(M_1'')$. Therefore, $\delta(M_3^{(n)}) \leq \delta(M_1'')$ for all n , since $\delta(M_3^{(0)}) \leq \delta(M_1'')$, and Φ_3 is a nondecreasing function. Hence $M^{(n)} \leq M''$ for all n , and $M^{(n)}$ converges. Denote $M''' = \lim_{n \rightarrow \infty} M^{(n)}$. Then $M''' \leq M''$. The optimal cost using a controller based on M''' is $[M'''X(0), X(0)]$. But $M''' \geq M''$, since M'' yields the minimal cost $[M''X(0), X(0)]$. Hence, $M^{(n)}$ converges, and the limit is $\lim_{n \rightarrow \infty} M^{(n)} = M''$. \square

We need only iteratively solve for $\delta(M_1^{(n)})$, which reduces the number of parameters that need to converge. Let P be the solution to the steady-state Riccati equation

$$0 = A^*P + PA + Q - P\bar{B}\bar{A}P^{-1}\bar{B}^*P.$$

A better initial value for $\delta(M_1)$ can be found as follows. Take

$$M_1 = P + (A^*P + PA) / (\max_i \lambda_i)$$

$$M_2 = M_1\bar{B}\bar{A}^{-1} + 4^*M_1\bar{B}\bar{A}^{-2}$$

$$\delta(M_1) = \Lambda^{-1}\delta(R + M_2^*B + B^*M_2).$$

The computations can be accelerated by using an intermediate update for $\delta(M_1)$. Let

$$\delta(M_1^{(n+1/2)}) = \Lambda^{-1}\delta[R - (M_1^{(n)} - \delta(M_1^{(n)}))\bar{A}\delta(M_1^{(n)})^{-1} \cdot (M_1^{(n)} - \delta(M_1^{(n)})) + M_2^{(n)*}B + B^*M_2^{(n)}].$$

Then $M^{(n+1)}$ is obtained by solving the algebraic Riccati equation

$$0 = A^*M^{(n+1)} + M^{(n+1)}A + Q - M^{(n+1)}\bar{B}\bar{A}\delta(M_1^{(n+1/2)})^{-1}\bar{B}^*M^{(n+1)}.$$

$$K = \begin{bmatrix} -0.0562 & 0.4086 & 0.7610 & 2.278 & -0.0507 & 0.0479 & 0 & 0.0446 & 0.0083 \\ 0.0463 & 0.2132 & 0.4914 & 2.264 & -0.0422 & 0.0523 & 0.0143 & 0 & 0.0085 \\ 0.0162 & 0.0258 & 0.0756 & 0.2742 & -0.1450 & 0.1344 & 0.0017 & 0.0053 & 0 \end{bmatrix}.$$

V. EXAMPLE: THREE MASSES WITH SPRINGS AND DAMPERS

A system of three masses and a fixed object linked by springs and dampers is given by

$$\begin{aligned}\ddot{y}_1 &= 2(\dot{y}_2 - \dot{y}_1) + 10(y_2 - y_1) + u_1 \\ 10\ddot{y}_2 &= 2(\dot{y}_1 - \dot{y}_2) + 10(y_1 - y_2) + .5(\dot{y}_3 - \dot{y}_2) + (y_3 - y_2) + u_2 \\ 2\ddot{y}_3 &= .5(\dot{y}_2 - \dot{y}_3) + (y_2 - y_3) - 2\dot{y}_3 - 8y_3 + u_3\end{aligned}$$

where y_i is the displacement of mass i , and the control input u_i is a force on mass i . The augmented state vector is $X = [y_1, \dot{y}_1, y_2, \dot{y}_2, y_3, \dot{y}_3, u_1, u_2, u_3]^T$. The control signals u_1, u_2, u_3 are assigned to three asynchronous processors with rates $\lambda_1 = 15, \lambda_2 = 5, \lambda_3 = 3$. The optimal steady-state Poisson LQR controller for $Q = I, R = I$ is $v(t) = -KX(t)$ is found in the equation at the bottom of the previous page. The diagonal terms in the part of K that is multiplied by $u_i(t)$ are zero, hence the update value $v_i(t)$ does not depend on $u_i(t)$, but may depend on other components of $u(t)$.

If $u(0) = 0$, the cost is given by $[M_1 x(0), x(0)]$, and $Tr.M_1$ is a good figure of merit for our design. For the continuous time LQR controller, a good figure of merit is $Tr.P$, where P is the solution to the continuous time steady-state Riccati equation. For our system, $Tr.P = 35.88$ and $Tr.M_1 = 37.61$, an increase of less than 5%. The cost is increased due to sampling and uncertainty in the sampling times.

VI. CONCLUSION

We derived the optimal control for a Poisson LQR asynchronous multirate controller. The optimal control for the i th processor involved feedback of both the state and control signals from other processors. The optimal control gain can be found by an iterative algorithm.

REFERENCES

- [1] N. Amit and J. D. Powell, "Optimal control of multirate systems," in *Proc. AIAA Guidance Contr. Conf.*, Albuquerque, NM, Aug. 1981.
- [2] M. C. Berg, N. Amit, and J. Powell, "Multirate digital control system design," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 1139–1150, 1988.
- [3] K. Bichteler, J.-B. Graveriaux, and J. Jacod, *Malliavin Calculus for Processes with Jumps*. New York: Gordon and Breach, 1987.
- [4] P. Bremaud, *Point Processes and Queues: Martingale Dynamics*. New York: Springer-Verlag, 1981.
- [5] D. P. Glasson, "A new technique for multirate digital control design and sample rate selection," *AIAA J. Guid. Contr.*, vol. 5, pp. 379–382, 1982.
- [6] Y. Ji and H. J. Chizeck, "Jump linear quadratic Gaussian control in continuous time," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1884–1892, 1992.
- [7] R. E. Kalman and J. E. Bertram, "A unified approach to the theory of sampling systems," *J. Franklin Institute*, no. 267, pp. 405–436, 1959.
- [8] N. N. Krasovskii and E. A. Lidskii, "Analytical design of controllers in systems with random attributes I, II, III," *Automation Remote Contr.*, vol. 22, pp. 1021–1025, 1141–1146, and 1289–1294, 1961. New York: Academic, 1967.
- [9] J. P. Lepeltier and B. Marchal, "Theorie generale du controle impulsional Markovien," *SIAM J. Contr. Optimiz.*, vol. 22, pp. 645–665, 1984.
- [10] J. L. Menaldi and M. Robin, "On singular stochastic control problems for diffusion with jumps," *IEEE Trans. Automat. Contr.*, vol. 29, pp. 991–1004, 1984.
- [11] M. Marton, *Jump Linear Systems in Automatic Control*. New York: Marcel Dekker, 1990.
- [12] D. D. Sworwer, "Feedback control of a class of linear systems with jump parameters," *IEEE Trans. Automat. Contr.*, vol. 14, pp. 9–14, 1969.
- [13] —, "Control of jump parameter systems with discontinuous state trajectories," *IEEE Trans. Automat. Contr.*, vol. 17, pp. 740–741, 1972.
- [14] D. D. Sworwer and V. G. Robinson, "Feedback regulations for jump parameter systems with state and control dependent transition rates," *IEEE Trans. Automat. Contr.*, vol. 18, pp. 355–360, 1973.
- [15] H. K. Wimmer, "Monotonicity of maximal solutions of algebraic Riccati equations," *Syst. Contr. Lett.*, vol. 5, pp. 317–319, 1985.
- [16] W. M. Wonham, "Random differential equations in control theory," in *Probabilistic Methods in Applied Mathematics*, A. T. Bharucha-Reid, Ed. New York: Academic, 1970, pp. 131–213.

On Solving Diophantine Equations by Real Matrix Manipulation

Manabu Yamada, Piao Chung Zun, and Yasuyuki Funahashi

Abstract—This note presents simple algorithms for obtaining the solutions of the Diophantine equation. Our methods can produce classes of all solutions with lower degree than a specified number. The previous algorithms involve some troublesome computations, e.g., the calculation of both the controllability indexes and the observability indexes or the solution of a pole assignment problem, etc. Our contribution is that our algorithm requires only basic matrix operations such as addition, subtraction, multiplication, and inversion of given real matrices. In addition, by solving simple linear equations, the class of all minimum degree solutions can be given. Therefore the computational efforts are reduced compared with previous algorithms.

I. INTRODUCTION

Consider the following problem: given three polynomial matrices, $D(s) \in \mathbb{R}[s]^{p \times p}$, $N(s) \in \mathbb{R}[s]^{p \times m}$, and $L(s) \in \mathbb{R}[s]^{p \times l}$, find polynomial matrices, $X(s) \in \mathbb{R}[s]^{p \times l}$ and $Y(s) \in \mathbb{R}[s]^{m \times l}$, such that

$$D(s)X(s) + N(s)Y(s) = L(s) \quad (1)$$

where $\mathbb{R}[s]^{p \times m}$ denotes a set of $p \times m$ polynomial matrices with real coefficients. We assume that $D(s)$ and $N(s)$ are left coprime and, without loss of generality, $D(s)$ is row-reduced [2, p. 68]. It is well-known that the former assumption ensures existence of solutions to (1) for arbitrary $L(s)$ [1]–[3]. Equation (1) is termed the Diophantine equation and plays an important role in many different aspects of linear system theory. Our interest is to develop a simple algorithm for obtaining solutions of the aforementioned problem, to which considerable attention has been paid during the last years [4]–[9].

The algorithms of [4]–[7] present the class of all solutions. Those of [4] and [9] provide the minimal degree solutions in terms of the rows of $Y(s)$ and the columns of $[X(s)^T, Y(s)^T]^T$, respectively, and left coprimeness of $D(s)$ and $N(s)$ is not required. However, their algorithms involve some troublesome computational efforts: the procedure in [4] requires controllability indexes and observability indexes of a realization of $D(s)^{-1}N(s)$ and the division of a polynomial matrix. The size of the matrix in [9], in which linearly

Manuscript received April 12, 1993; revised January 12, 1994 and July 25, 1994.

M. Yamada is with the Research Center for Micro-Structure Devices, Nagoya Institute of Technology, Nagaya 466, Japan.

P. C. Zun is with the Department of Automatic Control, North East Heavy Machinery Institute, Qiqihar, China.

Y. Funahashi is with the Department of Mechanical Engineering, Nagoya Institute of Technology, Nagoya 466, Japan.

IEEE Log Number 9407022.

dependent columns are sought by a column-searching algorithm, is very large. The algorithms in [5] and [6] involve a solution of a pole assignment problem and determination of a real number so as to make a matrix nonsingular. The synthesis of [7] requires selection of some appropriate feedback matrices to make cyclic the realization of $D(s)^{-1}N(s)$ and calculation of inversion of a polynomial matrix.

This note presents a simple algorithm for obtaining the class of all solutions with lower degree than a specified number in terms of $Y(s)$. For the computational aspects, we have some advantages. The proposed procedure requires no aforementioned troublesome computation. Our algorithm is performed by only basic matrix operations such as addition, subtraction, multiplication, and inversion of given real matrices. In addition, by solving simple linear equations, the class of all solutions with the minimum degree in terms of the columns of $Y(s)$ can be also given. Thus, the computational efforts are reduced compared with previous algorithms.

II. PRELIMINARIES

Let $\delta_{r,i}[\cdot]$ denote the i th row degree, i.e., the highest degree of all entries of the i th row vector, of a polynomial matrix and let $\delta_{c,i}[\cdot]$ denote the i th column degree. Assume that $\delta_{c,i}[D(s)] > \delta_{c,i}[N(s)]$, $i = 1 \cdots p$. This assumption assures that $D(s)^{-1}N(s)$ is strictly proper [1, p. 385]. This will be removed in the Appendix in a similar way to [7]. Define

$$d_i = \delta_{c,i}[D(s)], \quad i = 1 \cdots p$$

$$S(s) = \text{diag}\{s^{d_1}, i = 1 \cdots p\}, \quad (2)$$

$$\Psi(s) = \text{block diag}\{[s^{d_i-1} \cdots s-1], i = 1 \cdots p\}, \quad (3)$$

and let $D_h \in \mathbb{R}^{n \times p}$ and $D_l \in \mathbb{R}^{n \times p}$ be the highest row degree coefficient matrix and the lower row degree coefficient one of $D(s)$, respectively. From the aforementioned assumption, $D(s)$ and $N(s)$ are expressed by

$$D(s) = S(s)D_h + \Psi(s)D_l, \quad (4)$$

$$N(s) = \Psi(s)N_l \quad (5)$$

where $N_l \in \mathbb{R}^{m \times m}$. When all entries of $L(s)$ are separated into polynomials with the lower row degree than that of $S(s)$ and the remaining ones, the lower degree polynomial matrix can be represented by $\Psi(s)L_l$ and the remaining one by $S(s)L_h(s)$, i.e.,

$$L(s) = S(s)L_h(s) + \Psi(s)L_l, \quad (6)$$

where $L_l \in \mathbb{R}^{n \times l}$ and $L_h(s) \in \mathbb{R}[s]^{n \times l}$. Let

$$J = \text{block diag}\{J_1, J_2, \cdots, J_p\}$$

$$J_i = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{bmatrix} \in \mathbb{R}^{d_i \times d_i}, \quad i = 1 \cdots p$$

$$C_0 = \text{block diag}\{[1 \ 0 \ \cdots \ 0] \in \mathbb{R}^{d_i}, i = 1 \cdots p\}.$$

Since the row-reducedness of $D(s)$ means that D_h is nonsingular [2, p. 70], $D(s)^{-1}N(s)$ can be expressed by the following observer-form realization $\{A, B, C\}$ [1, p. 413]:

$$D(s)^{-1}N(s) = C(sI - A)^{-1}B \quad (7)$$

where

$$A = J - D_l D_h^{-1} C_0 \quad (8)$$

$$B = N_l \quad (9)$$

$$C = D_h^{-1} C_0. \quad (10)$$

III. MAIN RESULT

Lemma: Let

$$S^T(s)\bar{X}(s) + \Psi(s)\bar{Y}(s) = 0.$$

Then the class of all polynomial solutions $\bar{X}(s) \in \mathbb{R}[s]^l$, $\bar{Y}(s) \in \mathbb{R}[s]^{n \times l}$ of (11) are given by

$$\bar{X}(s) = C_0 Q(s), \quad (12)$$

$$\bar{Y}(s) = -(sI_n - J)Q(s) \quad (13)$$

where $Q(s) \in \mathbb{R}[s]^{n \times l}$ is an arbitrary polynomial matrix.

Proof: It can be easily checked that $(S(s), \Psi(s))$ is left coprime, $(C_0, sI_n - J)$ is right coprime, and

$$S(s)C_0 = \Psi(s)(sI_n - J).$$

Therefore (12) and (13) hold ([2, p. 186] and [3, p. 46]). \square

Let $\delta[\cdot]$ denote the highest degree of all entries of a polynomial matrix. Define

$$F(s) = L_l - D_l D_h^{-1} L_h(s),$$

$$f = \delta F(s).$$

Theorem: Given an integer $r \geq \max(f, n)$, define

$$W_r = [B \ AB \ \cdots \ A^r B] \in \mathbb{R}^{n \times m(r+1)}$$

$$F(s) = \sum_{i=0}^r F_i s^i \in \mathbb{R}[s]^{n \times l} \quad (14)$$

where $F_i \in \mathbb{R}^{n \times l}$. Then the class of all polynomial solutions $X(s) \in \mathbb{R}[s]^{n \times l}$ and $Y(s) \in \mathbb{R}[s]^{m \times l}$ of (1) with $\delta Y(s) \leq r$ are given by

$$X(s) = D_h^{-1} L_h(s) + U(s) + R(s)Z, \quad (15)$$

$$Y(s) = V(s) + P(s)Z \quad (16)$$

where $Z \in \mathbb{R}^{m(r+1) \times l}$ is an arbitrary real matrix and

$$U(s) = \sum_{i=0}^{r-1} U_i s^i \in \mathbb{R}[s]^{n \times l},$$

$$V(s) = \sum_{i=0}^r V_i s^i \in \mathbb{R}[s]^{m \times l},$$

$$R(s) = \sum_{i=0}^{r-1} R_i s^i \in \mathbb{R}[s]^{n \times m(r+1)},$$

$$P(s) = \sum_{i=0}^r P_i s^i \in \mathbb{R}[s]^{m \times m(r+1)}.$$

$$U_j = (A^j B)^T (W_r W_r^T)^{-1} \left(\sum_{i=0}^r A^i F_i \right) \in \mathbb{R}^{n \times l}, \quad j = 0 \cdots r$$

$$P_j = \left[\underbrace{0}_{m \times j} \ I_m \ 0 \right] - (A^j B)^T (W_r W_r^T)^{-1} W_r \in \mathbb{R}^{m \times m(r+1)},$$

$$U_j = C \left(\sum_{i=0}^{r-j-1} A^i F_{i+j+1} - \sum_{i=0}^{r-j-1} A^i B V_{i+j+1} \right) \in \mathbb{R}^{n \times l},$$

$$R_j = -C \sum_{i=0}^{r-j-1} A^i B P_{i+j+1} \in \mathbb{R}^{n \times m(r+1)}, \quad j = 0 \cdots r-1.$$

Proof: Substituting (4), (5), and (6) into (1) leads to

$$\begin{aligned} D(s)X(s) + N(s)Y(s) - L(s) \\ = S(s)\{D_h X(s) - L_h(s)\} \\ + \Psi(s)\{D_l X(s) + B Y(s) - L_l\} \\ = 0. \end{aligned} \quad (17)$$

From the Lemma, it follows that

$$\begin{aligned} X(s) &= D_h^{-1}\{L_h(s) + C_0 Q(s)\} \\ &= D_h^{-1} L_h(s) + C' Q(s) \end{aligned} \quad (18)$$

$$(sI_n - A)Q(s) + B Y(s) = F(s). \quad (19)$$

Letting

$$Q(s) = \sum_{i=0}^{r-1} Q_i s^i, \quad (20)$$

$$Y(s) = \sum_{i=0}^{r-1} Y_i s^i, \quad (21)$$

and substituting into (19), it follows that

$$\left. \begin{aligned} Q_{r-1} + B Y_r &= 0 \\ Q_{r-2} + B Y_{r-1} - A Q_{r-1} &= 0 \\ \vdots & \\ Q_{j-1} + B Y_j - A Q_j &= F_j \\ \vdots & \\ Q_0 + B Y_1 - A Q_1 &= F_1 \end{aligned} \right\} \quad (22)$$

$$B Y_0 - A Q_0 = F_0. \quad (23)$$

From (22)

$$\begin{aligned} Q_j &= \sum_{i=0}^{j-1} A^i F_{i+j+1} - \sum_{i=0}^{j-1} A^i B Y_{i+j+1} \in \mathbb{R}^{n \times l}, \\ 0 \leq j &\leq r-1. \end{aligned} \quad (24)$$

Substituting into (23) gives

$$W_r [Y_0^T \ Y_1^T \ \cdots \ Y_{r-1}^T]^T = \sum_{i=0}^{r-1} A^i F_i. \quad (25)$$

Since (A, B, C) is minimal [1, p. 439, Theorem 6.5-1], $W_r W_r^T$ is nonsingular. Hence, the general solution is given by

$$\begin{aligned} [Y_0^T \ Y_1^T \ \cdots \ Y_{r-1}^T]^T \\ = W_r^T (W_r W_r^T)^{-1} \sum_{i=0}^{r-1} A^i F_i + \{I_{m(r+1)} - W_r^T (W_r W_r^T)^{-1} W_r\} Z \\ = [Y_0^T \ Y_1^T \ \cdots \ Y_{r-1}^T]^T + [P_0^T \ P_1^T \ \cdots \ P_{r-1}^T]^T Z \end{aligned} \quad (26)$$

where $Z \in \mathbb{R}^{m(r+1) \times l}$ is arbitrary. Therefore, (16) holds. Substituting (24) and (26) into (18) gives (15). The proof is completed. \square

Remark: The theorem presents a simple algorithm for obtaining the class of all solutions with lower degree than a specified number $r+1$ in terms of $Y(s)$. The contribution of this algorithm is that the solutions can be computed easily by only basic matrix operations such as inversion of $W_r W_r^T$, whose size is $n \times n$, independent of a specified number r .

For a matrix M , let $[M]_j$ and $[M]_-$ be the j th column vector of M and a matrix satisfying

$$M[M]_- = M \quad (27)$$

respectively. $[M]_-$ is termed as the generalized inverse matrix of M . The solutions $\{X(s), Y(s)\}$ whose $\delta_i Y(s)$, $i = 1 \cdots l$ are minimum among all solutions of (1) are called the solutions with the minimum column degree in terms of $Y(s)$. The following corollary provides both the minimum possible number of r in the Theorem and the class of all solutions with the minimum column degree in terms of $Y(s)$.

Corollary: Let $f_j = \delta_{r_j} F(s)$ and

$$r_j^* = \min \left\{ k \geq 0 : \text{rank} \left[W_k, \sum_{i=0}^{l_j} A^i [F_i]_{i,j} \right] = \text{rank } W_k \right\}, \quad j = 1 \cdots l. \quad (28)$$

Given a set of integers $\{r_j, j = 1 \cdots l\}$ such that $r_j \geq r_j^*$, the class of all polynomial solutions $X(s) \in \mathbb{R}[s]^{n \times l}$ and $Y(s) \in \mathbb{R}[s]^{m \times l}$ of (1) with $\delta_i Y(s) \leq r_j, j = 1 \cdots l$ are given by

$$[X(s)]_{i,j} = D_h^{-1} [L_h(s)]_{i,j} + u_j^*(s) + R_j^*(s) z_j, \quad j = 1 \cdots l \quad (29)$$

$$[Y(s)]_{i,j} = v_j^*(s) + P_j^*(s) z_j, \quad j = 1 \cdots l \quad (30)$$

where $z_j \in \mathbb{R}^{m(r_j+1)}$, $j = 1 \cdots l$ are arbitrary real vectors and

$$u_j^*(s) = \sum_{i=0}^{\max(l_j, r_j)-1} u_{j,i}^* s^i \in \mathbb{R}[s]^n,$$

$$v_j^*(s) = \sum_{i=0}^{r_j} v_{j,i}^* s^i \in \mathbb{R}[s]^m,$$

$$R_j^*(s) = \sum_{i=0}^{\max(l_j, r_j)-1} R_{j,i}^* s^i \in \mathbb{R}[s]^{n \times m(r_j+1)},$$

$$P_j^*(s) = \sum_{i=0}^{r_j} P_{j,i}^* s^i \in \mathbb{R}[s]^{m \times m(r_j+1)}.$$

$$[(v_{j0}^*)^T (v_{j1}^*)^T \cdots (v_{j,r_j}^*)^T]^T = [W_{r_j}]_- \left(\sum_{i=0}^{l_j} A^i [F_i]_{i,j} \right) \in \mathbb{R}^n,$$

$$[(P_{j0}^*)^T (P_{j1}^*)^T \cdots (P_{j,r_j}^*)^T]^T = I_{m(r_j+1)} - [W_{r_j}]_- W_{r_j}^T \in \mathbb{R}^{m(r_j+1) \times m(r_j+1)}.$$

$$u_{j,i}^* = C' \left(\sum_{k=0}^{l_j-i-1} A^k [F_{i+k+1}]_{i,j} - \sum_{k=0}^{r_j-i-1} A^k B v_{j,i+k+1}^* \right) \in \mathbb{R}^n, \quad i = 0 \cdots \max(f_j, r_j) - 1,$$

$$R_{j,i}^* = -C' \sum_{k=0}^{l_j-i-1} A^k B P_{j,i+k+1}^* \in \mathbb{R}^{n \times m(r_j+1)}, \quad i = 0 \cdots \max(f_j, r_j) - 1.$$

With the choice of $\{r_j = r_j^*, j = 1 \cdots l\}$, (29) and (30) provide the class of all solutions with the minimum column degree in terms of $Y(s)$.

Proof. Without loss of generality, we consider the equations of the j th column in (18) and (19), i.e.,

$$[X(s)]_{i,j} = D_h^{-1} [L_h(s)]_{i,j} + C' [Q(s)]_{i,j}, \quad (31)$$

$$(sI_n - A)[Q(s)]_{i,j} + B[Y(s)]_{i,j} = [F(s)]_{i,j}. \quad (32)$$

Letting

$$[Q(s)]_{rj} = \sum_{i=0}^{\max(f_j, r_j)-1} q_{ji} s^i$$

$$[Y(s)]_{rj} = \sum_{i=0}^{r_j} y_{ji} s^i,$$

and substituting into (32), it follows that

$$q_{ji} = \sum_{k=0}^{f_j-i-1} A^k [F_{i+k+1}]_{rj} - \sum_{k=0}^{r_j-i-1} A^k B y_{j, i+k+1} \in \mathbf{R}^p,$$

$$i = 0 \cdots \max(f_j, r_j) - 1 \quad (33)$$

and

$$W_{rj} [y_{j0}^T \ y_{j1}^T \ \cdots \ y_{j, r_j}^T]^T = \sum_{i=0}^{f_j} A^i [F_i]_{rj}. \quad (34)$$

There exists a $[y_{j0}^T y_{j1}^T \cdots y_{j, r_j}^T]^T \in \mathbf{R}^{m(r_j+1)}$ satisfying (34) for a given number r_j if and only if

$$\text{rank} \left[W_{rj}, \sum_{i=0}^{f_j} A^i [F_i]_{rj} \right] = \text{rank} [W_{rj}]$$

[10, p. 29]. Then the class of all solutions satisfying (34) is given by

$$[y_{j0}^T \ y_{j1}^T \ \cdots \ y_{j, r_j}^T]^T$$

$$= [W_{rj}]^{-1} \left\{ \sum_{i=0}^{f_j} A^i [F_i]_{rj} \right\} + \{I_{m(r_j+1)} - [W_{rj}]^{-1} W_{rj}\} z_j$$

$$= [(y_{j0}^*)^T (y_{j1}^*)^T \cdots (y_{j, r_j}^*)^T]^T$$

$$+ [(P_{j0}^*)^T (P_{j1}^*)^T \cdots (P_{j, r_j}^*)^T]^T z_j \quad (35)$$

where $z_j \in \mathbf{R}^{m(r_j+1)}$, $j = 1 \cdots l$ are arbitrary real vectors. Hence, (30) holds. Substituting (33) and (35) into (31) leads to (29). \square

Remark: In [4] and [9], algorithms for the minimum degree solutions are proposed. The algorithm of [4] presents a solution in which the rows of $Y(s)$ has minimum degree. That of [9] results in a solution in which the columns of $[X(s)^T \ Y(s)^T]^T$ has minimal degree. Therefore, our algorithm proposes the minimum degree solution in a different sense from those of [4] and [9].

IV. NUMERICAL EXAMPLES

A. Example 1

Consider the following polynomial matrices.

$$D(s) = \begin{bmatrix} s+1 & s+2 \\ 0 & s+3 \end{bmatrix}, \quad N(s) = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix},$$

$$L(s) = I_2.$$

The observer-form realization of $D(s)^{-1}N(s)$ is given by

$$A = \begin{bmatrix} -1 & -1 \\ 0 & -3 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}.$$

We choose $r = n - 1 = 1$. Then

$$W = \begin{bmatrix} 0 & 1 & -1 & -2 \\ 1 & 1 & -3 & -3 \end{bmatrix}.$$

From the Theorem, the general solutions of (1) under the constraint of $\delta X(s) = 0$ and $\delta Y(s) \leq 1$ are shown in the equation found at the bottom of the page, where $Z \in \mathbf{R}^{1 \times 2}$ is arbitrary.

B. Example 2

In this example, we compare our algorithm with that of [9] from the viewpoint of the minimum degree solution.

Consider the following polynomial matrices.

$$D(s) = \begin{bmatrix} s^2 + s + 1 & s \\ s^2 & 2 \end{bmatrix}, \quad N(s) = \begin{bmatrix} s+1 & s^2 \\ s & s+1 \end{bmatrix},$$

$$L(s) = \begin{bmatrix} s^3 + 1 & s \\ 3s + 2 & s^2 + 1 \end{bmatrix}.$$

Thus, $D(s)$ and $N(s)$ are the same as those of example in [9]. By using the column searching algorithm of [9], the minimum degree solution is obtained as follows:

$$X(s) = \begin{bmatrix} 1 & 3 \\ s+1 & 4s+3 \end{bmatrix}, \quad Y(s) = \begin{bmatrix} -2s & -2s-3 \\ s & -5 \end{bmatrix}.$$

Second, our algorithm is applied. By our Corollary, we have $r_1^* = 0$ and $r_2^* = 1$. Then the class of all solutions with the minimum column degree in terms of $Y(s)$ is obtained as follows:

$$X(s) = \begin{bmatrix} -2 & -0.3 - 3z \\ s^2 - 2s - 1 & 1.1s^2 + 0.7s + 0.8 + (s^2 - 3s - 2)z \end{bmatrix},$$

$$Y(s) = \begin{bmatrix} 3 & (2s+3)(0.1+z) \\ 4 & -1.1s - 0.6 - (s-4)z \end{bmatrix},$$

where $z \in \mathbf{R}$ is arbitrary.

Therefore, it is seen that the algorithm of [9] can produce the solution with lower column degree in terms of $[X(s)^T \ Y(s)^T]^T$ than ours and, on the other hand, ours can produce the class of all lower column degree solutions in terms of $Y(s)$ than that of [9].

V. CONCLUSION

This note presents two results, the Theorem and Corollary. Theorem shows a simple algorithm for obtaining the class of all solutions with lower degree than a specified number $r+1$ in terms of $Y(s)$, where r is an arbitrary integer $r \geq \max(f, n)$. The contribution is that the class of solutions can be computed easily by only basic matrix operations such as inversion of W, W_r^T , whose size is $n \times n$, independent of a specified number r . Corollary provides the minimum possible number of r in the Theorem. So the class of all solutions with the minimum column degree in terms of $Y(s)$ can be also obtained. The necessary computation is a generalized inverse matrix of W_{rj} in the Corollary, that is simple linear equations of (27).

In also [4] and [9], simple methods are proposed. However, the procedure in [4] needs controllability indexes and observability indexes of a realization of $D(s)^{-1}N(s)$ and the division of a polynomial matrix. In [9], it is required to find l linearly dependent

$$X(s) = \begin{bmatrix} 0.5 & -0.4 \\ 0 & 0.3 \end{bmatrix} + \begin{bmatrix} 0.4 & -0.1 & 0.3 & -0.2 \\ -0.3 & -0.3 & -0.1 & -0.1 \end{bmatrix} Z$$

$$Y(s) = \begin{bmatrix} 0.5s - 0.5 & -0.4s + 0.3 \\ -0.5s + 0.5 & 0.1s - 0.2 \end{bmatrix} + \begin{bmatrix} 0.4s + 0.7 & -0.1s + 0.2 & 0.3s + 0.4 & -0.2s - 0.1 \\ -0.1s + 0.2 & 0.4s + 0.7 & -0.2s - 0.1 & 0.3s + 0.4 \end{bmatrix} Z$$

columns in a real matrix, whose size is $(N + K + 1)p \times \{(K + 1)p + (K + 2)l\}$, where

$$N = \max \{\delta D(s), \delta L(s)\},$$

$$K = \max [n, \delta L(s), \delta \{N(s) \text{adj } D(s)\}] \geq n.$$

Thus, the computational efforts of our algorithm are much reduced compared with the previous ones.

APPENDIX

In this Appendix, the assumption of $\delta_i D(s) > \delta_i N(s)$ $i = 1 \cdots p$, is removed. That is, we consider the case of $\delta_i D(s) \leq \delta_i N(s)$ for some i . In this case, $N(s)$ can be represented as follows [1]–[3]:

$$N(s) = D(s)Q_N(s) + R_N(s)$$

where $Q_N(s) \in \mathbb{R}[s]^{p \times m}$, $R_N(s) \in \mathbb{R}[s]^{p \times m}$, and $\delta_i D(s) > \delta_i R_N(s)$, $i = 1 \cdots p$. In particular, if $\delta_i D(s) \geq \delta_i N(s)$ for $i = 1 \cdots p$, i.e., $D(s)^{-1}N(s)$ is proper, then $N(s)$ can be represented by $N_h \in \mathbb{R}^{p \times m}$ and $N_l \in \mathbb{R}^{p \times m}$ as follows:

$$N(s) = S(s)N_h + \Psi(s)N_l.$$

Hence Q_N and $R_N(s)$ can be obtained easily as

$$Q_N = D_h^{-1}N_h,$$

$$R_N(s) = \Psi(s)(N_l - D_l D_h^{-1}N_h).$$

The theorem can provide polynomial solutions $X_N(s)$ and $Y_N(s)$ satisfying

$$D(s)X_N(s) + R_N(s)Y_N(s) = L(s).$$

Then, by using the $X_N(s)$ and $Y_N(s)$, polynomial solutions $X(s)$ and $Y(s)$ of (1) are given by

$$\begin{aligned} X(s) &= X_N(s)Q_N(s)Y_N(s), \\ Y(s) &= Y_N(s). \end{aligned}$$

This can be readily verified by substituting into (1) as follows:

$$\begin{aligned} D(s)X(s) + N(s)Y(s) &= D(s)\{X_N(s) - Q_N(s)Y_N(s)\} \\ &+ \{D(s)Q_N(s) + R_N(s)\}Y_N(s) \\ &= L(s). \end{aligned}$$

REFERENCES

- [1] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1980.
- [2] F. M. Callier and C. A. Desoer, *Multivariable Feedback Systems*. New York: Springer-Verlag, 1982.
- [3] V. Kučera, *Discrete Linear Control*. New York: Wiley, 1979.
- [4] W. A. Wolovich and P. J. Antsaklis, "The canonical Diophantine equations with applications," *SIAM J. Contr. Optimiz.*, vol. 22, no. 5, pp. 777–787, 1984.
- [5] C. H. Fang and F. R. Chang, "A novel approach for solving Diophantine equations," *IEEE Trans. Circ. Syst.*, vol. 37, no. 11, pp. 1455–1457, 1990.
- [6] C. H. Fang, "A new method for solving the polynomial generalized Bezout identity," *IEEE Trans. Circ. Syst.-I: Fundamental Theory and Applications*, vol. 39, no. 1, pp. 63–65, 1992.
- [7] —, "A simple approach to solving the Diophantine equation," *IEEE Trans. Automat. Contr.*, vol. 37, no. 1, pp. 152–155, 1992.

- [8] J. Feinstein and Y. Bar-Ness, "The solution of the matrix polynomial equation $A(s)X(s) + B(s)Y(s) = C(s)$," *IEEE Trans. Automat. Contr.*, vol. 29, no. 1, pp. 75–77, 1984.
- [9] Y. S. Lai, "An algorithm for solving the matrix polynomial equation $B(s)D(s) + A(s)N(s) = H(s)$," *IEEE Trans. Circ. Syst.*, vol. 36, no. 8, pp. 1087–1089, 1989.
- [10] C. T. Chen, *Linear System Theory and Design*. New York: Rinehart and Winston, 1984.

Nonlinear Versus Linear Control in the Absolute Stabilizability of Uncertain Systems with Structured Uncertainty

Andrey V. Savkin and Ian R. Petersen

Abstract—This note considers a stabilization problem for a class of uncertain linear systems containing structured uncertainty described by a certain Integral Quadratic Constraint. The notion of stabilizability considered is that of absolute stabilizability. The main result gives a necessary and sufficient condition for the absolute stabilizability of this class of uncertain systems in terms of the existence of a solution to a corresponding "diagonally scaled" H^∞ control problem. It follows from this result that absolute stabilizability via nonlinear control implies absolute stabilizability via linear control.

1. INTRODUCTION

In recent years, a number of results have appeared showing that certain classes of uncertain linear systems have the property that if they can be stabilized via a nonlinear controller, then they can also be stabilized via a linear controller; see [1]–[4]. These results are concerned with problems of robust stabilization in which the underlying uncertain system contains "unstructured" uncertainty. This note establishes a corresponding result for a class of uncertain linear system containing structured uncertainty. The uncertain systems under consideration allow for nonlinear, time-varying, dynamic, noncausal uncertainties. This class of uncertainties is described by an "Integral Quadratic Constraint" such as found in the work of Yakubovich; e.g., see [5], [6]. The notion of stability considered is that of absolute stability which can also be found in the work of Yakubovich; e.g., see [5], [6]. Our main result is a necessary and sufficient condition for the absolute stabilizability of such an uncertain system. This condition is given in terms of the existence of a solution to a corresponding "diagonally scaled" H^∞ control problem. A corollary to this result is the fact that if an uncertain system is absolutely stabilizable via nonlinear (but time-invariant) control, then it will be absolutely stabilizable via linear control.

The launching point for this note is a recent result developed in [7]. This result is a convexity result for a collection of integral quadratic forms defined over the space of solutions to a stable linear time-invariant system; a related result can also be found in [8]. The result of [7] is referred to as an "S-procedure." In this note, we extend the S-procedure result of [7] to a collection of integral functionals defined over the space of solutions to a nonlinear time-invariant system. This

Manuscript received April 9, 1993; revised January 28, 1994. This work was supported by the Australia Research Council.

The authors are with the Department of Electrical Engineering, Australian Defence Force Academy, Campbell, 2600, Australia.
IEEE Log Number 9407024.

enables us to consider the absolute stabilization problem for the case in which nonlinear controllers are allowed. We can summarize our approach as a three-step procedure for robust stabilization based on the following three steps:

- the S-procedure theorem;
- H^∞ control;
- the Kalman–Yakubovich–Popov (KYP) lemma.

The remaining sections of the note proceed as follows. Section II introduces the class of uncertain systems under consideration and defines the notion of absolute stabilizability. Section III presents our “S-procedure” for nonlinear systems. Section IV presents our main results.

II. PROBLEM STATEMENT

We consider a class of uncertain linear systems described by state equations of the form

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t) + \sum_{s=0}^k D_s \xi_s(t) \\ y(t) &= Cx(t) + \sum_{s=0}^k H_s \xi_s(t) \\ z_s(t) &= K_s x(t) + G_s u(t); \quad s = 0, 1, \dots, k\end{aligned}\quad (2.1)$$

where $x(t) \in \mathbf{R}^n$ is the state, $u(t) \in \mathbf{R}^m$ is the control input, $y(t) \in \mathbf{R}^l$ is the measured output, $z_0(t) \in \mathbf{R}^{h_0}, \dots, z_k(t) \in \mathbf{R}^{h_k}$ are the uncertainty outputs, and $\xi_0(t) \in \mathbf{R}^{r_0}, \dots, \xi_k(t) \in \mathbf{R}^{r_k}$ are the uncertainty inputs.

System Uncertainty: The uncertainty in the above system is described by equations of the form

$$\begin{aligned}\xi_0(t) &= \phi_0(t, z_0(\cdot)|_0^t), \xi_1(t) = \phi_1(t, z_1(\cdot)|_0^t), \\ &\dots, \xi_k(t) = \phi_k(t, z_k(\cdot)|_0^t)\end{aligned}\quad (2.2)$$

where the following Integral Quadratic Constraint is satisfied. This uncertainty description allows for nonlinear, time-varying, dynamic, noncasual uncertainties.

Definition 2.1: (Integral Quadratic Constraint; see [5], [6].) An uncertainty of the form (2.2) is an admissible uncertainty for the system (2.1) if the following conditions hold. Given any locally square integrable control input $u(\cdot)$ and any corresponding solution to (2.1), (2.2) defined on an existence interval¹ $(0, t_*)$, then there exists a sequence $\{t_i\}_{i=1}^\infty$ and constants $d_0 \geq 0, d_1 \geq 0, \dots, d_k \geq 0$ such that $t_i \rightarrow t_*, t_i \geq 0$ and

$$\int_0^{t_i} \|\xi_s(t)\|^2 dt \leq \int_0^{t_i} \|z_s(t)\|^2 dt + d_s, \quad \forall t_i \text{ and } \forall s = 0, 1, \dots, k. \quad (2.3)$$

Here $\|\cdot\|$ denotes the standard Euclidean norm. Also, note that t_* and t_i may be equal to infinity.

We will consider the problem of absolute stabilization via a nonlinear output feedback controller

$$\dot{x}_c(t) = \Lambda(x_c(t), y(t)); \quad u(t) = \lambda(x_c(t), y(t)) \quad (2.4)$$

where $\Lambda(\cdot, \cdot)$ and $\lambda(\cdot, \cdot)$ are continuous vector functions.

Notation: Throughout this note, $L_2[0, \infty)$ denotes the Hilbert space of square integrable vector valued functions defined on $[0, \infty)$. Also $\|\cdot\|_2$ denotes the corresponding $L_2[0, \infty)$ norm.

¹That is, t_* is the upper limit of the time interval over which the solution exists.

Definition 2.2: The uncertain system (2.1), (2.2) is absolutely stabilizable via nonlinear control if there exists a controller of the form (2.4) and a constant $c > 0$ such that

- for any initial condition $[x(0) \ x_c(0)] = [x_0 \ x_{c0}]$ and any admissible $\xi(\cdot) = [\xi_0(\cdot) \ \dots \ \xi_k(\cdot)] \in L_2[0, \infty)$, the closed-loop system defined by (2.1) and (2.4) has a unique solution which remains bounded on $[0, \infty)$;
- the closed-loop system (2.1), (2.4) with $\xi(t) \equiv 0$ is globally uniformly asymptotically stable;
- for any initial condition $[x(0) \ x_c(0)] = [x_0 \ x_{c0}]$ and any admissible $\xi(\cdot) = [\xi_0(\cdot) \ \dots \ \xi_k(\cdot)]$, then $[x(\cdot) \ x_c(\cdot) \ u(\cdot) \ \xi_0(\cdot) \ \dots \ \xi_k(\cdot)] \in L_2[0, \infty)$, (hence, $t_* = \infty$) and

$$\begin{aligned}\|x(\cdot)\|_2^2 + \|x_c(\cdot)\|_2^2 + \|u(\cdot)\|_2^2 + \sum_{s=0}^k \|\xi_s(\cdot)\|_2^2 \\ \leq c \left[\|x(0)\|^2 + \|x_c(0)\|^2 + \sum_{s=0}^k d_s \right];\end{aligned}\quad (2.5)$$

- for any $\epsilon > 0$, there exists a $\delta > 0$ such that the following condition holds: for any vector $[x_0' \ x_{c0}']' \in \{h: \|h\| \leq \delta\}$ and any $\xi(\cdot) = [\xi_0(\cdot) \ \xi_1(\cdot) \ \dots \ \xi_k(\cdot)] \in L_2[0, \infty)$, let $[x^{(1)}(t)' \ x_c^{(1)}(t)']'$ be the corresponding solution to the closed-loop system defined by (2.1) and (2.3) with initial condition $x^{(1)}(0) = x_0, x_c^{(1)}(0) = x_{c0}$. Also let $[x^{(2)}(t)' \ x_c^{(2)}(t)']'$ denote the solution to the closed-loop system with initial condition $x^{(2)}(0) = 0, x_c^{(2)}(0) = 0$; then $\|x^{(1)}(\cdot) - x^{(2)}(\cdot)\|_2^2 + \|x_c^{(1)}(\cdot) - x_c^{(2)}(\cdot)\|_2^2 + \|u^{(1)}(\cdot) - u^{(2)}(\cdot)\|_2^2 < \epsilon$.

Note, condition iv) in the above definition is a technical stability condition which is needed in order to establish the results of this note.

Observation 2.1: If the uncertain system (2.1), (2.3) is absolutely stabilizable, then the corresponding closed-loop uncertain system (2.1), (2.3), (2.4) will have the property that $x(t) \rightarrow 0$ as $t \rightarrow \infty$ for any admissible uncertainty $\xi(\cdot)$. Indeed, since $\{x(\cdot), u(\cdot), \xi(\cdot)\} \in L_2[0, \infty)$, we can conclude from (2.1) that $\dot{x}(\cdot) \in L_2[0, \infty)$. However, using the fact that $x(\cdot) \in L_2[0, \infty)$ and $\dot{x}(\cdot) \in L_2[0, \infty)$, it now follows that $x(t) \rightarrow 0$ as $t \rightarrow \infty$. Furthermore, if the controller (2.4) is linear, then also $x_c(t) \rightarrow 0$ as $t \rightarrow \infty$.

III. S-PROCEDURE FOR NONLINEAR SYSTEMS

In this section, we present a result which extends the “S-Procedure” of [7]. The main result of this section applies to a nonlinear, time-invariant system of the form

$$\dot{h}(t) = \Pi(h(t), v(t)) \quad (3.1)$$

where $h(t) \in \mathbf{R}^n$ is the state and $v(t) \in \mathbf{R}^p$ is the input. Associated with the system (3.1) is the following set of functionals: $f_0(h(\cdot), v(\cdot)) = \int_0^\infty \nu_0(h(t), v(t)) dt, \dots, f_k(h(\cdot), v(\cdot)) = \int_0^\infty \nu_k(h(t), v(t)) dt$.

Assumptions: The system (3.1) and associated set of functionals satisfy the following assumptions.

- 1) The function $\Pi(\cdot, \cdot)$ is continuous.
- 2) For all $\{h(\cdot), v(\cdot)\} \in L_2[0, \infty)$, the quantities $f_0(h(\cdot), v(\cdot)), \dots, f_k(h(\cdot), v(\cdot))$ are finite.
- 3) Given any $\epsilon > 0$, there exists a constant $\delta > 0$ such that the following condition holds. For any input function $v_0(\cdot) \in L_2[0, \infty)$ and any $h_0 \in \{h_0 \in \mathbf{R}^n: \|h_0\| \leq \delta\}$, let $h_1(t)$ denote the corresponding solution to (3.1) with initial condition $h(0) = h_0$ and let $h_2(t)$ denote the corresponding solution to (3.1) with initial condition $h(0) = 0$. Then

both $h_1(t)$ and $h_2(t)$ are defined on $[0, \infty)$ and both functions belong to $L_2[0, \infty)$. Furthermore, $|f_s(h_1(\cdot), \psi(\cdot)) - f_s(h_2(\cdot), \psi(\cdot))| < \epsilon$ for $s = 0, 1, \dots, k$.

Note, Assumption 3.3) is a stability-type assumption on the system (3.1).

Notation: For the system (3.1) satisfying the above assumptions, we define $\Omega \subset L_2[0, \infty)$ as follows. Ω is the set of $\{h(\cdot), \psi(\cdot)\}$ such that $\psi(\cdot) \in L_2[0, \infty)$ and $h(\cdot)$ is the corresponding solution to (3.1) with $h(0) = 0$.

Theorem 3.1: Consider the system (3.1) and associated functionals and suppose the Assumptions 3.1)–3.3) are satisfied. If $f_0(h(\cdot), \psi(\cdot)) \geq 0$ for all $\{h(\cdot), \psi(\cdot)\} \in \Omega$ such that $f_1(h(\cdot), \psi(\cdot)) \geq 0, \dots, f_k(h(\cdot), \psi(\cdot)) \geq 0$, then there exist constants $\tau_0 \geq 0, \tau_1 \geq 0, \dots, \tau_k \geq 0$ such that $\sum_{s=0}^k \tau_s > 0$ and

$$\tau_0 f_0(h(\cdot), \psi(\cdot)) \geq \tau_1 f_1(h(\cdot), \psi(\cdot)) + \tau_2 f_2(h(\cdot), \psi(\cdot)) + \dots + \tau_k f_k(h(\cdot), \psi(\cdot)) \quad (3.2)$$

for all $\{h(\cdot), \psi(\cdot)\} \in \Omega$.

In order to prove this theorem, we will use the following convex analysis result (the proof of which was given in the preliminary version of this note [9]). However, we first introduce some notation.

Notation: Given $S \subset \mathbb{R}^n$ and $T \subset \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, then $S + T := \{x + y: x \in S, y \in T\}$, $\lambda S := \{\lambda x: x \in S\}$, and $\text{cone}(S) := \{\alpha x: x \in S, \alpha \geq 0\}$. Also, $\text{cl}(S)$ denotes the closure of the set S .

Lemma 3.1: (See [9] for proof.) Consider a set $M \subset \mathbb{R}^{k+1}$ with the property that $a + b \in \text{cl}(M)$ for all $a, b \in M$. If $x_0 \geq 0$ for all vectors $[x_0 \dots x_k] \in M$ such that $x_1 \geq 0, \dots, x_k \geq 0$, then there exist constants $\tau_0 \geq 0, \dots, \tau_k \geq 0$ such that $\sum_{s=0}^k \tau_s > 0$ and $\tau_0 x_0 \geq \tau_1 x_1 + \tau_2 x_2 + \dots + \tau_k x_k$ for all $[x_0 \dots x_k] \in M$.

Proof of Theorem 3.1: Suppose $f_0(h(\cdot), \psi(\cdot)) \geq 0$ for all $\{h(\cdot), \psi(\cdot)\} \in \Omega$ such that $f_1(h(\cdot), \psi(\cdot)) \geq 0, \dots, f_k(h(\cdot), \psi(\cdot)) \geq 0$ and let $M := \{[f_0(h(\cdot), \psi(\cdot)) \dots f_k(h(\cdot), \psi(\cdot))] \in \mathbb{R}^{k+1}: \{h(\cdot), \psi(\cdot)\} \in \Omega\}$. It follows from the assumption on the set Ω that $x_0 \geq 0$ for all vectors $[x_0 \dots x_k] \in M$ such that $x_1 \geq 0, \dots, x_k \geq 0$. Now let $\{h_a(\cdot), \psi_a(\cdot)\} \in \Omega$ and $\{h_b(\cdot), \psi_b(\cdot)\} \in \Omega$ be given. Since $h_a(\cdot) \in L_2[0, \infty)$, there exists a sequence $\{T_i\}_{i=1}^\infty \subset \mathbb{R}$ such that $T_i \rightarrow \infty$ and $h_a(T_i) \rightarrow 0$. Now consider the corresponding sequence $\{h_i(\cdot), \psi_i(\cdot)\}_{i=1}^\infty \subset \Omega$, where

$$\psi_i(t) = \begin{cases} \psi_a(t) & t < T_i \\ \psi_b(t - T_i) & t \geq T_i. \end{cases}$$

We will establish that $f_s(h_i(\cdot), \psi_i(\cdot)) \rightarrow f_s(h_a(\cdot), \psi_a(\cdot)) + f_s(h_b(\cdot), \psi_b(\cdot))$ for $s = 0, 1, \dots, k$. Indeed, let $s \in \{0, 1, \dots, k\}$ be given and fix i . Now suppose $\hat{h}_i(\cdot)$ is the solution to (3.1) with input $\psi(\cdot) = \psi_b(\cdot)$ and $h(0) = h_a(T_i)$. It follows from the time invariance of the system (3.1) that $h_i(t) \equiv \hat{h}_i(t - T_i)$. Hence,

$$\begin{aligned} f_s(h_i(\cdot), \psi_i(\cdot)) &= \int_0^\infty \nu_s(h_i(t), \psi_i(t)) dt \\ &= \int_0^{T_i} \nu_s(h_a(t), \psi_a(t)) dt \\ &\quad + \int_{T_i}^\infty \nu_s(h_i(t), \psi_b(t - T_i)) dt \\ &= \int_0^{T_i} \nu_s(h_a(t), \psi_a(t)) dt + f_s(\hat{h}_i(t), \psi_b(t)). \end{aligned}$$

Using the fact that $h_a(T_i) \rightarrow 0$, Assumption 3.3 implies $f_s(\hat{h}_i(t), \psi_b(t)) \rightarrow f_s(h_b(\cdot), \psi_b(\cdot))$ as $i \rightarrow \infty$. Also, $\int_0^{T_i} \nu_s(h_a(t), \psi_a(t)) dt \rightarrow \int_0^\infty \nu_s(h_a(t), \psi_a(t)) dt \rightarrow f_s(h_a(\cdot), \psi_a(\cdot))$. Hence, $f_s(h_i(\cdot), \psi_i(\cdot)) \rightarrow f_s(h_a(\cdot), \psi_a(\cdot)) + f_s(h_b(\cdot), \psi_b(\cdot))$.

From the above, it follows that the set M has the property that $a + b \in \text{cl}(M)$ for all $a, b \in M$. Hence, Lemma 3.1 implies that there exist constants $\tau_0 \geq 0, \dots, \tau_k \geq 0$ such that $\sum_{s=0}^k \tau_s > 0$ and $\tau_0 x_0 \geq \tau_1 x_1 + \dots + \tau_k x_k$ for all $[x_0 \dots x_k] \in M$. That is, condition (3.2) is satisfied. \square

Note a feature of the proof of this result is to exploit the “fading memory” property of a nonlinear system satisfying Assumptions 3.1)–3.3).

IV. THE MAIN RESULT

Our main results require that the uncertain system (2.1), (2.3) satisfies the following assumptions.

Assumptions. Let matrices \hat{D} , \hat{K} , E_1 , and \hat{E}_2 be defined by $\hat{D} := [D_0 \dots D_k]$, $\hat{K} := [K'_0 \dots K'_k]'$, $\hat{E}_1 = \sum_{s=0}^k G'_s G_s$, $\hat{E}_2 = \sum_{s=0}^k H_s H'_s$. Then, we assume

- 4.1) The pair (A, B) is stabilizable.
- 4.2) The pair (A, K) is observable.
- 4.3) $E_1 > 0$.
- 4.4) The pair (A, C) is detectable.
- 4.5) The pair (A, D) is controllable.
- 4.6) $\hat{E}_2 > 0$.

Assumptions 4.1) and 4.4) are necessary for the “nominal system” to be stabilizable and hence can be made without loss of generality. Assumptions 4.2), 4.3), 4.5), and 4.6) are technical assumptions required to ensure that the underlying H^∞ problem is “nonsingular;” e.g., see [10]. This underlying “diagonally scaled” H^∞ control problem is defined as follows. Let $\tau_1 > 0, \dots, \tau_k > 0$ be given constants and consider the system

$$\dot{x}(t) = Ax(t) + Du(t) + Bw(t)$$

$$y(t) = Cx(t) + Dw(t)$$

$$z(t) = Kx(t) + Gu(t) \quad (4.1)$$

where $u(t)$ is the disturbance input, $z(t)$ is the controlled output,

$$K := \begin{bmatrix} K_0 \\ \sqrt{\tau_1} K_1 \\ \vdots \\ \sqrt{\tau_k} K_k \end{bmatrix}, \quad G := \begin{bmatrix} G_0 \\ \sqrt{\tau_1} G_1 \\ \vdots \\ \sqrt{\tau_k} G_k \end{bmatrix}.$$

$$H := [H_0 \quad \frac{1}{\sqrt{\tau_1}} H_1 \quad \dots \quad \frac{1}{\sqrt{\tau_k}} H_k],$$

$$D := [D_0 \quad \frac{1}{\sqrt{\tau_1}} D_1 \quad \dots \quad \frac{1}{\sqrt{\tau_k}} D_k]. \quad (4.2)$$

Also, consider the following induced norm bound condition:

$$J \triangleq \sup_{u(\cdot) \in L_2[0, \infty), \|u(\cdot)\|_2 = 1} \frac{\|z(\cdot)\|_2^2}{\|w(\cdot)\|_2^2} < 1. \quad (4.3)$$

The H^∞ control problem corresponding to the system (4.1) is said to have a solution if there exists a controller mapping from y to u such that the closed-loop system satisfies condition (4.3); e.g., see [10].

Theorem 4.1: Consider the uncertain system (2.1), (2.3) and suppose that Assumptions 4.1)–4.6) are satisfied. Then the following statements are equivalent.

- i) The uncertain system (2.1), (2.3) is absolutely stabilizable via a nonlinear controller of the form (2.4).

- ii) There exist constants $\tau_1 > 0, \dots, \tau_k > 0$ such that the H^∞ control problem (4.1), (4.3) is solved by a linear controller of the form

$$\dot{x}_c(t) = A_c x_c(t) + B_c y(t); \quad u(t) = C_c x_c(t). \quad (4.4)$$

If condition ii) holds, then the uncertain system (2.1), (2.3) is absolutely stabilizable via the linear controller (4.4).

In order to prove this result, we first establish the following preliminary lemma.

Lemma 4.1: Suppose that the uncertain system (2.1), (2.3) is absolutely stabilizable via a controller of the form (2.4). Then, there exist constants $\tau_1 > 0, \dots, \tau_k > 0$ and $\delta > 0$ such that

$$\begin{aligned} & \|z_0(\cdot)\|_2^2 - \|\xi_0(\cdot)\|_2^2 + \sum_{s=1}^k \tau_s (\|z_s(\cdot)\|_2^2 - \|\xi_s(\cdot)\|_2^2) \\ & \leq -\delta \left(\|x(\cdot)\|_2^2 + \|x_c(\cdot)\|_2^2 + \|u(\cdot)\|_2^2 + \sum_{s=0}^k \|\xi_s(\cdot)\|_2^2 \right) \end{aligned} \quad (4.5)$$

for all the solutions to the closed-loop system with the initial condition $x(0) = 0, x_c(0) = 0$.

Proof: Suppose that the uncertain system (2.1), (2.3) is absolutely stabilizable via a control of the form (2.4) and consider the corresponding closed-loop system. This system is described by the state equations

$$\begin{aligned} \dot{x}(t) &= A x(t) B \lambda(x_c(t), y(t)) + \sum_{s=0}^k D_s \xi_s(t) \\ \dot{x}_c(t) &= \Lambda(x_c(t), y(t)). \end{aligned} \quad (4.6)$$

This system can be regarded as a system of the form (3.1) with the associations: $h(t) \sim [x(t)' \ x_c(t)']'$ and $v(t) \sim \xi(t)$ where $\xi(\cdot) = [\xi_0(\cdot) \ \dots \ \xi_k(\cdot)]$. The corresponding set $\Omega \subset L_2[0, \infty)$ defined in Section III is as follows: Ω is the set of vector functions $\{x(\cdot), x_c(\cdot), \xi(\cdot)\}$ such that $\xi(\cdot) \in L_2[0, \infty)$ and $[x(\cdot)' \ x_c(\cdot)']'$ is the corresponding solution to (2.1), (2.4) with initial condition $x(0) = 0, x_c(0) = 0$. Also associated with the above system is the following set of integral functionals $f_0(\cdot), \dots, f_k(\cdot)$:

$$\begin{aligned} f_0(x(\cdot), x_c(\cdot), \xi(\cdot)) &:= \int_0^\infty (\|\xi_0(t)\|^2 - \|K_0 x(t) + G_0 \lambda(x_c(t), y(t))\|^2) dt \\ &\quad - \delta_0 \int_0^\infty (\|x(t)\|^2 + \|x_c(t)\|^2 \\ &\quad + \|\lambda(x_c(t), y(t))\|^2 + \|\xi(t)\|^2) dt \\ f_1(x(\cdot), x_c(\cdot), \xi(\cdot)) &:= \int_0^\infty (\|K_1 x(t) + G_1 \lambda(x_c(t), y(t))\|^2 - \|\xi_1(t)\|^2) dt \\ &\quad + \delta_0 \int_0^\infty (\|x(t)\|^2 + \|x_c(t)\|^2 \\ &\quad + \|\lambda(x_c(t), y(t))\|^2 + \|\xi(t)\|^2) dt \\ &\vdots \\ f_k(x(\cdot), x_c(\cdot), \xi(\cdot)) &:= \int_0^\infty (\|K_k x(t) + G_k \lambda(x_c(t), y(t))\|^2 - \|\xi_k(t)\|^2) dt \\ &\quad + \delta_0 (\|x(t)\|^2 + \|x_c(t)\|^2 \\ &\quad + \|\lambda(x_c(t), y(t))\|^2 + \|\xi(t)\|^2) dt \end{aligned} \quad (4.7)$$

where $\delta_0 = 1/2(k+1)c$ and c is as defined in (2.1). Definition 2.2. Thus, using (2.1), we can write

$$\begin{aligned} f_0(x(\cdot), x_c(\cdot), \xi(\cdot)) &= -(\|z_0(\cdot)\|_2^2 - \|\xi_0(\cdot)\|_2^2 \\ &\quad + \delta_0 (\|x(\cdot)\|_2^2 + \|x_c(\cdot)\|_2^2 \\ &\quad + \|\xi(\cdot)\|_2^2 + \|u(\cdot)\|_2^2)) \end{aligned}$$

$$\begin{aligned} f_1(x(\cdot), x_c(\cdot), \xi(\cdot)) &= \|z_1(\cdot)\|_2^2 - \|\xi_1(\cdot)\|_2^2 + \delta_0 (\|x(\cdot)\|_2^2 \\ &\quad + \|x_c(\cdot)\|_2^2 + \|\xi(\cdot)\|_2^2 + \|u(\cdot)\|_2^2) \end{aligned}$$

\vdots

$$\begin{aligned} f_k(x(\cdot), x_c(\cdot), \xi(\cdot)) &= \|z_k(\cdot)\|_2^2 - \|\xi_k(\cdot)\|_2^2 + \delta_0 (\|x(\cdot)\|_2^2 \\ &\quad + \|x_c(\cdot)\|_2^2 + \|\xi(\cdot)\|_2^2 + \|u(\cdot)\|_2^2). \end{aligned} \quad (4.8)$$

We wish to apply Theorem 3.1 to the system (4.6) and associated functionals (4.7). However, we must first show that Assumptions 3.1)–3.3) are satisfied. The satisfaction of Assumption 3.1 follows directly from the continuity of the functions $\Lambda(\cdot, \cdot)$ and $\lambda(\cdot, \cdot)$. Assumption 3.2 follows directly from part iii) of Definition 2.2 and equations (4.8). To establish Assumption 3.3, let $\epsilon > 0$ be given and consider an uncertainty input $\xi(\cdot) \in L_2[0, \infty)$. Let $[x^{(1)}(t)' \ x_c^{(1)}(t)']'$ be the corresponding solution to the closed-loop system (4.6) with $[x^{(1)}(0) \ x_c^{(1)}(0)]' = [x_0 \ x_{c0}]$, and let $[x^{(2)}(t)' \ x_c^{(2)}(t)']'$ be the solution to (4.6) with $x^{(2)}(0) = 0, x_c^{(2)}(0) = 0$. Now consider the quantity $|f_0(x^{(1)}(\cdot), x_c^{(1)}(\cdot), \xi(\cdot)) - f_0(x^{(2)}(\cdot), x_c^{(2)}(\cdot), \xi(\cdot))|$ where the functional $f_0(\cdot, \cdot, \cdot)$ is defined in (4.7). Using the triangle inequality and the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned} & |f_0(x^{(1)}(\cdot), x_c^{(1)}(\cdot), \xi(\cdot)) - f_0(x^{(2)}(\cdot), x_c^{(2)}(\cdot), \xi(\cdot))| \\ & \leq \delta_0 (\|x^{(1)}(\cdot) - x^{(2)}(\cdot)\|_2^2 + \|x_c^{(1)}(\cdot) - x_c^{(2)}(\cdot)\|_2^2 \\ & \quad + \|u^{(1)}(\cdot) - u^{(2)}(\cdot)\|_2^2 + \|z_0^{(1)}(\cdot) - z_0^{(2)}(\cdot)\|_2^2) \\ & \leq \delta_0 (\|x^{(1)}(\cdot) - x^{(2)}(\cdot)\|_2^2 \\ & \quad + \|x_c^{(1)}(\cdot) - x_c^{(2)}(\cdot)\|_2^2 + \|u^{(1)}(\cdot) - u^{(2)}(\cdot)\|_2^2) \\ & \quad + \|[K_0 \ D_0]\|^2 (\|x^{(1)}(\cdot) - x^{(2)}(\cdot)\|_2^2 \\ & \quad + \|u^{(1)}(\cdot) - u^{(2)}(\cdot)\|_2^2) \end{aligned} \quad (4.9)$$

where $\|[K_0 \ D_0]\|$ denotes the induced matrix norm of the matrix $[K_0 \ D_0]$. However, it follows from part iv) of Definition 2.2 that given any $\epsilon > 0$, there exists a constant $\delta > 0$ such that $\|x_0\|^2 + \|x_{c0}\|^2 \leq \delta^2$ implies $\|x^{(1)}(\cdot) - x^{(2)}(\cdot)\|_2^2 + \|x_c^{(1)}(\cdot) - x_c^{(2)}(\cdot)\|_2^2 + \|u^{(1)}(\cdot) - u^{(2)}(\cdot)\|_2^2 < \epsilon$. Now using inequality (4.9), it follows that $|f_0(x^{(1)}(\cdot), x_c^{(1)}(\cdot), \xi(\cdot)) - f_0(x^{(2)}(\cdot), x_c^{(2)}(\cdot), \xi(\cdot))| \leq (\delta_0 + \|[K_0 \ D_0]\|)\epsilon$. A similar inequality holds for the quantities $|f_1(x^{(1)}(\cdot), x_c^{(1)}(\cdot), \xi(\cdot)) - f_1(x^{(2)}(\cdot), x_c^{(2)}(\cdot), \xi(\cdot))|, \dots, |f_k(x^{(1)}(\cdot), x_c^{(1)}(\cdot), \xi(\cdot)) - f_k(x^{(2)}(\cdot), x_c^{(2)}(\cdot), \xi(\cdot))|$. Thus, by choosing

$$\epsilon = \min \left\{ \frac{\epsilon}{(\delta_0 + \|[K_s \ D_s]\|)} : s = 0, 1, \dots, k \right\}$$

and letting δ equal the corresponding value of δ , it follows that Assumption 3.3 is satisfied.

To apply Theorem 3.1, we must show that $f_0(x(\cdot), x_c(\cdot), \xi(\cdot)) \geq 0$ for all $\{x(\cdot), x_c(\cdot), \xi(\cdot)\} \in \Omega$ such that $f_s(x(\cdot), x_c(\cdot), \xi(\cdot)) \geq 0$ for $s = 1, \dots, k$. If this is not true, then there exists a $\{x^0(\cdot), x_c^0(\cdot), \xi^0(\cdot)\} \in \Omega$ such that $f_0(x^0(\cdot), x_c^0(\cdot), \xi^0(\cdot)) < 0$ and

$f_s(x^0(\cdot), x^0(\cdot), \xi^0(\cdot)) \geq 0$ for $s = 1, \dots, k$. However, the input $\xi^0(\cdot)$ satisfies the integral quadratic constraint (2.3) with $t_r = \infty$ and $d_s = \delta_0(\|x^0(\cdot)\|_2^2 + \|x^0(\cdot)\|_2^2 + \|\xi^0(\cdot)\|_2^2 + \|u^0(\cdot)\|_2^2)$ for $s = 0, 1, \dots, k$. Hence, condition iii) of Definition 2.2 implies that

$$\begin{aligned} & \|x^0(\cdot)\|_2^2 + \|x^0(\cdot)\|_2^2 + \|\xi^0(\cdot)\|_2^2 + \|u^0(\cdot)\|_2^2 \\ & \leq c \sum_{s=0}^k \delta_0(\|x^0(\cdot)\|_2^2 + \|x^0(\cdot)\|_2^2 + \|\xi^0(\cdot)\|_2^2 + \|u^0(\cdot)\|_2^2) \\ & \leq c(k+1)\delta_0(\|x^0(\cdot)\|_2^2 + \|x^0(\cdot)\|_2^2 + \|\xi^0(\cdot)\|_2^2 + \|u^0(\cdot)\|_2^2) \\ & \leq \frac{1}{2}(\|x^0(\cdot)\|_2^2 + \|x^0(\cdot)\|_2^2 + \|\xi^0(\cdot)\|_2^2 + \|u^0(\cdot)\|_2^2) \end{aligned}$$

which yields the desired contradiction.

We now apply Theorem 3.1 to the system (4.6) and functionals (4.7). It follows that there exist $\tau_0 \geq 0, \tau_1 \geq 0, \dots, \tau_k \geq 0$ such that $\sum_{s=0}^k \tau_s > 0$ and $\tau_0 f_0(x(\cdot), x(\cdot), \xi(\cdot)) \geq \sum_{s=1}^k \tau_s f_s(x(\cdot), x(\cdot), \xi(\cdot))$ for all $\{x(\cdot), x(\cdot), \xi(\cdot)\} \in \Omega$. We now prove that $\tau_s > 0$ for all $s = 0, 1, \dots, k$. If we let $\delta := \delta_0 \sum_{s=0}^k \tau_s > 0$, then

$$\begin{aligned} & \sum_{s=0}^k \tau_s (\|z_s(\cdot)\|_2^2 - \|\xi_s(\cdot)\|_2^2) \\ & \leq -\delta \left(\|x(\cdot)\|_2^2 + \|x(\cdot)\|_2^2 + \|u(\cdot)\|_2^2 + \sum_{s=0}^k \|\xi_s(\cdot)\|_2^2 \right) \quad (4.10) \end{aligned}$$

for all $\{x(\cdot), x(\cdot), \xi(\cdot)\} \in \Omega$. Now if $\tau_j = 0$ for some $j \in \{0, \dots, k\}$, then we can let $\xi_j(\cdot) \equiv 0$ for all $s \neq j$ and choose any nonzero $\xi_j(\cdot) \in L_2[0, \infty)$. However, this leads to a contradiction with (4.10) since the left side of (4.10) is nonnegative and the right side of (4.10) is negative. Therefore, we must have, $\tau_s > 0$ for $s = 0, 1, \dots, k$. Furthermore, observe that in this case, we may take $\tau_0 = 1$ without loss of generality. Moreover, with $\tau_0 = 1$, inequality (4.10) leads directly to (4.5). \square

Proof of Theorem 4.1: i) \Rightarrow ii) Consider an uncertain system (2.1), (2.3) satisfying Assumptions 4.1)–4.6) and suppose that the system is absolutely stabilizable via a nonlinear controller of the form (2.4). It follows from Lemma 4.1 that there exist constants $\tau_1 > 0, \tau_2 > 0, \dots, \tau_k > 0$ such that inequality (4.5) is satisfied. Now the system (2.1) is equivalent to the system (4.1) where $w(\cdot) = [\xi_0(\cdot), \sqrt{\tau_1}\xi_1(\cdot), \dots, \sqrt{\tau_k}\xi_k(\cdot)]$, $\hat{z}(\cdot) = [z_0(\cdot), \sqrt{\tau_1}z_1(\cdot), \dots, \sqrt{\tau_k}z_k(\cdot)]$, and the matrices D, K, G , and H are defined as in (4.2). It follows from inequality (4.5) that there exists a constant $\delta > 0$ such that if $x(0) = 0$, then $\|\hat{z}(\cdot)\|_2^2 - \|w(\cdot)\|_2^2 = \|z_0(\cdot)\|_2^2 + \sum_{s=1}^k \tau_s \|z_s(\cdot)\|_2^2 - \|\xi_0(\cdot)\|_2^2 - \sum_{s=1}^k \tau_s \|\xi_s(\cdot)\|_2^2 \leq -\delta \sum_{s=0}^k \|\xi_s(\cdot)\|_2^2$ for all $w(\cdot) \in L_2[0, \infty)$. Hence, there exists a $\delta_1 > 0$ such that if $x(0) = 0$, then $\|\hat{z}(\cdot)\|_2^2 - \|w(\cdot)\|_2^2 \leq -\delta_1(\|\xi_0(\cdot)\|_2^2 + \sum_{s=1}^k \tau_s \|\xi_s(\cdot)\|_2^2) = -\delta_1 \|w(\cdot)\|_2^2$ for all $w(\cdot) \in L_2[0, \infty)$. That is, condition (4.3) is satisfied. Therefore, the controller (2.4) solves the standard H^∞ control problem (4.1), (4.3). Furthermore, it follows from Assumptions 4.1)–4.6) that the system (4.1) satisfies the assumptions required by Theorems 5.5 and 5.6 of [10]. Hence, using this theorem, it follows that there exists a linear controller of the form (4.4) which solves the H^∞ control problem (4.1), (4.3).

ii) \Rightarrow i) Consider an uncertain system (2.1), (2.3) satisfying Assumptions 4.1)–4.6) and suppose there exist constants $\tau_1 > 0, \tau_2 > 0, \dots, \tau_k > 0$ such that the H^∞ control problem (4.1), (4.3) is solved by a linear controller of the form (4.4). The closed-loop system defined by (4.1) and (4.4) is described by the state equations

$$\dot{\bar{x}}(t) = P\bar{x}(t) + Qw(t); \quad \hat{z}(t) = T\bar{x}(t) \quad (4.11)$$

where

$$\bar{x} = \begin{bmatrix} x \\ x_r \end{bmatrix}, \quad P = \begin{bmatrix} A & BC_r \\ B_r C & A_r \end{bmatrix},$$

$$Q = \begin{bmatrix} D \\ D_r H \end{bmatrix}, \quad T = [K \quad GC_r].$$

Since the controller (4.4) solves the H^∞ control problem mentioned above, the matrix P will be stable; e.g., see [10]. Also, if we apply the control (4.4) to the uncertain system (2.1), (2.3), the resulting closed-loop uncertain system can be described by the state equations (4.11) where $w(t)$ is now interpreted as an uncertainty input and $\hat{z}(t)$ is interpreted as an uncertainty output.

We will consider an uncertain system described by the state equations (4.11) where the uncertainty is required to satisfy the following integral quadratic constraint analogous to (2.3):

$$\int_0^{t_1} \|T\bar{x}(t)\|^2 dt \leq \int_0^{t_1} \|w(t)\|^2 dt + d \quad \forall t_1. \quad (4.12)$$

Note that any admissible uncertainty for the closed-loop uncertain system (2.1), (2.3), (4.4) will also be an admissible uncertainty for the uncertain system (4.11), (4.12) with $d = d_0 + \sum_{s=1}^k \tau_s d_s \geq 0$.

We will now show that the uncertain system (4.11), (4.12) is absolutely stable; see [5], [6] for the definition of absolute stability. First, recall that the system (4.11) is such that (4.3) is satisfied. Hence, there exists a constant $\delta_1 > 0$ such that if $x(0) = 0$, then $\|\hat{z}(\cdot)\|_2^2 - \|w(\cdot)\|_2^2 \leq -\delta_1 \|w(\cdot)\|_2^2$ for all $w(\cdot) \in L_2[0, \infty)$. Therefore, if $x(0) = 0$, then $\int_0^\infty (\|T\bar{x}(t)\|^2 - \|w(t)\|^2) dt \leq -\delta_1 \int_0^\infty \|w(t)\|^2 dt$ for all $w(\cdot) \in L_2[0, \infty)$. Furthermore, the stability of the matrix P implies that there exists a constant $\mu > 0$ such that if $x(0) = 0$, then $\int_0^\infty \|x(t)\|^2 dt \leq \mu \int_0^\infty \|w(t)\|^2 dt$ for all $w(\cdot) \in L_2[0, \infty)$. Hence, the constant $\delta = \delta_1/(\mu + 1) > 0$ is such that

$$\int_0^\infty (\|T\bar{x}(t)\|^2 - \|w(t)\|^2) dt \leq -\delta \int_0^\infty (\|x(t)\|^2 + \|w(t)\|^2) dt \quad (4.13)$$

for all $\{x(\cdot), w(\cdot)\} \in L_2[0, \infty)$ connected by (4.11) with $x(0) = 0$. Using a version of the KYP Lemma (Theorem 1 of [6]), condition (4.13) and the stability of the matrix P now imply that the uncertain system (4.11), (4.12) is absolutely stable.

We will now show that the closed-loop uncertain system (2.1), (2.3), (4.4) satisfies the conditions for absolute stabilizability given in Definition 2.2. Condition i) of Definition 2.2 follows directly from the absolute stability of the uncertain system (4.11), (4.12). Moreover, the absolute stability of this uncertain system implies that there exists a constant $c_0 > 0$ such that for any initial condition $x(0) = x_0$ and any admissible uncertainty input $w(\cdot)$ described by (4.12), then $\{x(\cdot), w(\cdot)\} \in L_2[0, \infty)$ and $\|x(\cdot)\|_2^2 + \|w(\cdot)\|_2^2 \leq c_0[\|x_0\|^2 + d]$. Now equation (4.4) implies that there exists a constant $c_1 > 0$ such that $\|u(t)\| \leq c_1 \|x(t)\|$ for all solutions to the closed-loop system (2.1), (4.4). Furthermore, given any admissible $\xi(\cdot) = [\xi_0(\cdot) \dots \xi_k(\cdot)]$ for the closed-loop system (2.1), (2.3), (4.4), then $w(\cdot) = [\xi_0(\cdot), \sqrt{\tau_1}\xi_1(\cdot), \dots, \sqrt{\tau_k}\xi_k(\cdot)]$ is an admissible uncertainty input for the uncertain system (4.11), (4.12) with $d = d_0 + \sum_{s=1}^k \tau_s d_s$. Hence, we can conclude that

$$\begin{aligned} & \frac{1}{2c_1} \|u(\cdot)\|_2^2 + \frac{1}{2} \|x(\cdot)\|_2^2 + \|\xi_0(\cdot)\|_2^2 + \sum_{s=1}^k \tau_s \|\xi_s(\cdot)\|_2^2 \\ & \leq \|x(\cdot)\|_2^2 + \|\xi_0(\cdot)\|_2^2 + \sum_{s=1}^k \tau_s \|\xi_s(\cdot)\|_2^2 \\ & \leq c_0 \left[\|x_0\|^2 + d_0 + \sum_{s=1}^k \tau_s d_s \right]. \quad (4.14) \end{aligned}$$

However, it is straightforward to verify that there exist constants $\sigma_1 > 0$ and $\sigma_2 > 0$ such that we can write $\sigma_1(\|u(\cdot)\|_2^2 + \|x(\cdot)\|_2^2 + \sum_{s=0}^k \|\xi_s(\cdot)\|_2^2) \leq (1/2\sigma_1)\|u(\cdot)\|_2^2 + \frac{1}{2}\|x(\cdot)\|_2^2 + \|\xi_0(\cdot)\|_2^2 + \sum_{s=1}^k \tau_s \|\xi_s(\cdot)\|_2^2$, and moreover, $c_0[\|x_0\|^2 + d_0 + \sum_{s=1}^k \tau_s d_s] \leq \sigma_2[\|x_0\|^2 + \sum_{s=0}^k d_s]$. Hence, by combining these inequalities with (4.14), it follows that $\|u(\cdot)\|_2^2 + \|x(\cdot)\|_2^2 + \sum_{s=0}^k \|\xi_s(\cdot)\|_2^2 \leq \sigma_2/\sigma_1[\|x_0\|^2 + \sum_{s=0}^k d_s]$. That is, condition iii) of Definition 2.2 is satisfied. Conditions ii) and iv) of Definition 2.2 follows directly from the stability of the matrix P . Thus, we can now conclude that the system (2.1), (2.3) is absolutely stabilizable via the control (4.4). \square

Remark: The above theorem gives a necessary and sufficient condition for absolute stabilizability in terms of the solution to a corresponding H^∞ control problem. It is well known that such an H^∞ control problem can be solved in terms of two algebraic Riccati equations; e.g., see [10].

The following corollary is an immediate consequence of the above theorem.

Corollary 4.1: If the uncertain system (2.1), (2.3) satisfies Assumptions 4.1)–4.6) and is absolutely stabilizable via nonlinear control, then it will be absolutely stabilizable via a linear controller of the form (4.4).

REFERENCES

- [1] M. A. Rotea and P. P. Khargonekar, "Stabilization of uncertain systems with norm bounded uncertainty – A control Lyapunov approach," *SIAM J. Contr. Optimiz.*, vol. 27, no. 6, pp. 1462–1476, 1989.
- [2] P. P. Khargonekar and K. R. Poolla, "Uniformly optimal control of linear time-invariant plants: Nonlinear time-varying controllers," *Syst. Contr. Lett.*, vol. 6, no. 5, pp. 303–309, 1986.
- [3] P. P. Khargonekar, T. T. Georgiou, and A. M. Pascoal, "On the robust stabilizability of linear time invariant plants with unstructured uncertainty," *IEEE Trans. Automat. Contr.*, vol. AC-32, pp. 201–207, 1987.
- [4] K. R. Poolla and T. Ting, "Nonlinear time-varying controllers for robust stabilization," *IEEE Trans. Automat. Contr.*, vol. AC-32, pp. 195–200, 1987.
- [5] V. A. Yakubovich, "Dichotomy and absolute stability of nonlinear systems with periodically nonstationary linear part," *Syst. Contr. Lett.*, vol. 11, no. 3, pp. 221–228, 1988.
- [6] —, "Absolute stability of nonlinear systems with a periodically nonstationary linear part," *Sov. Phys. Doklady*, vol. 32, no. 1, pp. 5–7, 1988.
- [7] A. Megretsky and S. Treil, "Power distribution inequalities in optimization and robustness of uncertain systems," *J. Math. Syst. Estim. Contr.*, vol. 3, no. 3, pp. 301–319, 1993.
- [8] J. S. Shamma, "Robustness analysis for time-varying systems," in *Proc. 31st IEEE Conf. Dec. Contr.*, Tucson, AZ., Dec. 1992.
- [9] A. V. Savkin and I. R. Petersen, "Nonlinear versus linear control in the absolute stabilizability of uncertain linear systems with structured uncertainty," in *Proc. 32 IEEE Conf. Dec. Contr.*, San Antonio, TX., Dec. 1993.
- [10] T. Basar and P. Bernhard, *H^∞ -Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Boston: Birkhauser, 1991.

A Linear Algebraic Framework for Dynamic Feedback Linearization

E. Aranda-Bricaire, C. H. Moog, and J.-B. Pomet

Abstract—To any accessible nonlinear system we associate a set of infinitesimal Brunovsky form. This gives an algebraic criterion for strong accessibility as well as a generalization of Kronecker controllability indices. An output function which defines a right-invertible system without zero dynamics is shown to exist if and only if the basis of the Brunovsky form can be transformed into a system of exact differential forms. This is equivalent to the system being differentially flat and hence constitutes a necessary and sufficient condition for dynamic feedback linearizability.

1. INTRODUCTION

The problem of exact linearization of a nonlinear system using static state feedback was solved in [18] and [22]. It can be shown that this problem is linked with the classification of functions (or exact one-forms) with respect to their relative degree [8]. When the linearization problem can not be solved using static state feedback, it is appealing to try to solve it using dynamic state feedback. This note emphasizes the links between this new problem and the classification of (non necessarily exact) one-forms with respect to their relative degree. The dynamic feedback linearization problem was stated in its full generality for the first time in [6]: given a nonlinear control system

$$\Sigma: \dot{x} = f(x) + g(x)u \quad (1)$$

where $x \in R^n$, $u \in R^m$, find a dynamic compensator

$$C: \begin{cases} \dot{u} = \alpha(x, \xi) + \beta(x, \xi)v \\ \xi = \gamma(x, \xi) + \delta(x, \xi)v \end{cases} \quad (2)$$

and an extended set of coordinates $z = \alpha(x, \xi)$ in which the extended system reads as a controllable linear one. Relying upon the differential geometric approach, sufficient conditions and necessary conditions have been given in [7]. In some particular cases, necessary and sufficient conditions are given there. A less general formulation of the dynamic linearization problem is as follows. Consider a nonlinear system where the output $y = h(x)$, $y \in R^m$, has been specified. If the system is right-invertible, it is always possible to construct a dynamic compensator in such a way that noninteracting control is achieved [10], [26]. The standard decoupling feedback provides also input–output linearization and if, in addition, the system has the property of having no zero dynamics, then it actually solves the dynamic linearization problem [19], [21]. Thus, the existence of such an output function is a sufficient condition for dynamic linearizability.

In [12], [13], and [24], the notions of linearizing output and endogeneous feedback were introduced. A linearizing output is a system of functions φ_i of x, u, \dot{u}, \dots which are differentially

Manuscript received February 5, 1993; revised September 15, 1993 and March 1, 1994. The work of E. Aranda-Bricaire was supported by CONACYT and CINVESTAV-IPN, Mexico.

E. Aranda-Bricaire and C. H. Moog are with the Laboratoire d'Automatique de Nantes (URA CNRS 823), Ecole Centrale de Nantes – Université de Nantes, 1 rue de la Noë, 44072 Nantes cedex 03, France.

J.-B. Pomet is with INRIA, B. P. 93, 06902 Sophia-Antipolis cedex, France. Part of this work was performed while J.-B. Pomet was with the Laboratoire d'Automatique de Nantes.

IEEE Log Number 9407025.

independent and such that both x and u can be expressed as functions of the φ 's and their time derivatives. Existence of a linearizing output is called differential flatness of the system. A dynamic compensator of the form (2) such that ξ can be expressed as a function of x, u, \dot{u}, \dots , is called endogeneous. flatness is equivalent to dynamic linearizability by endogeneous feedback. Furthermore, recent results [14], [15] show that no system is linearizable by nonendogeneous feedback without being flat. An equivalence relation between systems may be defined such that differential flatness is equivalence to a linear system. This equivalence corresponds to transformations by endogeneous feedback.

In [23], the notion of a dynamic equivalence has been defined in terms of D -algebras. A system is equivalent to a controllable linear system if and only if its associated D -algebra is free. The set of generators of this D -algebra plays the role of linearizing output.

The goal of the present note is to characterize the existence of these linearizing outputs depending on x, u, \dot{u}, \dots , defined in a new algebraic approach which can be recovered from the early notions of zero dynamics and infinite zero structure [9], [19], [20], [26]. This characterization is equivalent to the system's property of being differentially flat, and hence constitutes, from the results in [14] and [15], a necessary and sufficient condition for dynamic feedback linearizability.

It will be shown that to any nonlinear system, one can associate a so-called infinitesimal Brunovsky form which may be viewed as a time-varying Brunovsky form of the first-order approximation of Σ [11]. The construction of this Brunovsky form provides an accessibility criterion, as well as a generalization of linear Kronecker indices. This form singles out a family of m elements of the formal vector space of differential forms, and it is shown that a linearizing output exists if, and only if, this family can be transformed into a system of exact one-forms via some invertible transformation. A preliminary version of this work was presented in [28], where the infinitesimal Brunovsky form was defined (it was called nonexact instead of infinitesimal), and was shown to provide a tool to characterize linearizing outputs in the sense of [12], [13], [24]. Let us also mention that a more "differential geometric" presentation of the present material may be found in [2] and [29].

Section II is devoted to some preliminaries from [9], and to a problem statement of dynamic linearization in terms of the infinite zero structure, as in [19]. The infinitesimal Brunovsky form is introduced in Section III with an algorithmic construction. An accessibility criterion is given which involves purely algebraic computations. In Section IV, existence of a linearizing output is characterized in terms of the infinitesimal Brunovsky form. The above theory is illustrated in Section V by the study of various particular cases. Concluding remarks are offered in Section VI.

II. PRELIMINARIES

A. The Infinite Zero Structure [9], [26]

Consider the nonlinear control system Σ , where $f(\cdot)$ and the columns of $g(\cdot)$ are meromorphic vector fields. Throughout the note it is assumed that $\text{rank } g(x) = m$. Let \mathcal{K} denote the field of meromorphic functions of x, u, \dot{u}, \dots . The time derivative of a function $\varphi \in \mathcal{K}$ is defined by

$$\frac{d}{dt}\varphi = \dot{\varphi} = \frac{\partial \varphi}{\partial x}(f(x) + g(x)u) + \sum_{j \geq 0} \frac{\partial \varphi}{\partial u^{(j)}} u^{(j+1)}.$$

Clearly, \mathcal{K} is closed under time-differentiation. Let \mathcal{E} denote the \mathcal{K} -vector space spanned by $dx, du, d\dot{u}, \dots$. The elements of \mathcal{E} are called

differential forms of degree one, or simply one-forms. d/dt induces a derivation on \mathcal{E} in the following way $d/dt: \omega = \sum_j a_j dv_j \mapsto \dot{\omega} = \sum_j (\dot{a}_j dv_j + a_j d\dot{v}_j)$. The relative degree of a one-form $\omega \in \text{span}_{\mathcal{K}}\{dx\}$ is defined as the smallest integer r such that $\omega^{(r)} \notin \text{span}_{\mathcal{K}}\{dx\}$. If such an integer does not exist, set $r = \infty$.

Now, consider the system Σ and suppose that the output function $y = h(x)$, $y \in \mathbb{R}^m$, has been specified. Introduce the chain of subspaces $\mathcal{E}_0 \subset \mathcal{E}_1 \subset \dots \subset \mathcal{E}_n$ of \mathcal{E} , defined by

$$\mathcal{E}_k := \text{span}_{\mathcal{K}}\{dx, dy, \dots, dy^{(k)}\}. \quad (3)$$

The number of zeros at infinity of order less than or equal to k , for $1 \leq k \leq n$, is

$$\sigma_k = \dim \frac{\mathcal{E}_k}{\mathcal{E}_{k-1}}. \quad (4)$$

The infinite zero structure can be given either by the list $\{\sigma_k\}$ or by the list $\{n'_i\}$ of the orders of the zeros at infinity. Σ is said to be (right) invertible if $\sigma_n = m$. Following [9], one has the following.

Lemma 2.1: Assume that Σ is invertible. Let $\mathcal{X} := \text{span}_{\mathcal{K}}\{dx\}$, $\mathcal{Y} := \text{span}_{\mathcal{K}}\{dy^{(k)}, k \geq 0\}$. Then $\dim(\mathcal{X} \cap \mathcal{Y}) = \Sigma, n'_i$.

Remark 2.2: Note that although \mathcal{Y} is, in general, infinite dimensional, the intersection of subspaces $\mathcal{X} \cap \mathcal{Y}$ is, at most, of dimension n . The subspace $\mathcal{X} \cap \mathcal{Y}$ has been first considered in [5] for studying minimality in dynamic decoupling.

B. Dynamic Feedback Linearization Problem Statement

Any nonlinear system with outputs and which is right-invertible can be fully linearized whenever it has no zero dynamics, in the sense of the dynamics of the reduced inverse system [20]. Thus, the absence of zero dynamics is a sufficient condition for dynamic feedback linearization [19], [21]. This is equivalent to $\Sigma, n'_i = n$. This yields (more precisely) the following.

Problem Statement 1: Given Σ , find, if possible, an m -dimensional output function $y = h(x)$ such that the system is right-invertible and $\Sigma, n'_i = n$. ■

Solvability of this problem is not necessary for dynamic feedback linearizability. A more general approach to the dynamic feedback linearization problem, which originates in [12], [13], [24], where it is instrumental to define differential flatness, consists of allowing the output function to explicitly depend on the input u as well as on a finite number, say, $\nu - 1$, of its time derivatives. From Lemma 2.1, this situation may be stated as the existence of an m -dimensional output function $y = h(x, u, \dots, u^{(\nu-1)})$ such that the system is right-invertible and $\dim(\mathcal{X}_\nu \cap \mathcal{Y}) = n + m\nu$, where $\mathcal{X}_\nu := \text{span}_{\mathcal{K}}\{dx, du, \dots, du^{(\nu-1)}\}$. For square invertible systems, one has $\mathcal{X}_\nu + \mathcal{Y} = \mathcal{X} + \mathcal{Y}$ and consequently $\dim(\mathcal{X}_\nu \cap \mathcal{Y}) = \dim(\mathcal{X} \cap \mathcal{Y}) + m\nu$. So, the more general problem is stated as follows.

Problem Statement 2: Given Σ , find, if possible, an integer ν and an m -dimensional output function $y = h(x, u, \dots, u^{(\nu-1)})$ such that the system is right-invertible and

$$\dim(\mathcal{X} \cap \mathcal{Y}) = n. \quad (5)$$

If such an output exists, it is called a *linearizing output*. ■

C. Differential Flatness

In [12], [13], and [24], the notions of linearizing output, differential flatness, endogeneous and exogeneous feedback were introduced. In [12] and [13], this is done within a differential algebraic framework, whereas in [24] the analytic case is also considered.

Roughly speaking, a linearizing output [12], [13] is a system of differentially independent functions φ , of x, u, \dot{u}, \dots , such that x, u

can in turn be expressed as functions of the φ 's and a finite number of their time derivatives. This definition is equivalent to the one in Problem Statement 2 because condition (5) means that one is able to recover the variables x, u as functions of the outputs y , and their time derivatives, and on the other hand, right invertibility ensures that the outputs y are differentially independent in the sense that they do not satisfy any differential equation independent of u . We refer the reader to [16] for an exhaustive discussion of differential flatness and its link with an equivalence relation between systems.

A system Σ is said to be differentially flat (or simply flat) if it admits a system of linearizing outputs. It is proved in [12], [13] and [24] that differential flatness for a system Σ is equivalent to linearizability via a dynamic compensator C which has the property of being endogeneous (this may be defined as the possibility to express the variables ξ as functions of $x, u, u^{(1)}, \dots$). Recent results [14], [15] show that flatness is in fact equivalent to dynamic linearizability without any restriction on the nature of compensators.

III. THE INFINITESIMAL BRUNOVSKY FORM

A. A Flag for the Differential Vector Space \mathcal{E}

We shall construct a sequence of subspaces of \mathcal{E} in the following manner. Define

$$\mathcal{H}_0 = \text{span}_{\mathcal{K}} \{dx, du\}$$

$$\mathcal{H}_k = \{ \omega \in \mathcal{H}_{k-1} \mid \omega \in \mathcal{H}_{k-1} \} \quad (6)$$

It is clear that $\mathcal{E} \supset \mathcal{H}_0 \supset \mathcal{H}_1 \supset \mathcal{H}_2 \supset \dots$ and that at the first step the above induction yields $\mathcal{H}_1 = \text{span}_{\mathcal{K}} \{dx\}$. Proposition 3.1 is a simple consequence of the construction. Feedback invariance comes from the fact that the relative degree of a one-form is obviously invariant under regular static state feedback. Existence of the integer l comes from the fact that each \mathcal{H}_k is a finite dimensional \mathcal{K} -vector space so that at each step either its dimension decreases or $\mathcal{H}_{k+1} = \mathcal{H}_k$.

Proposition 3.1 \mathcal{H}_k is the space of one-forms which have relative degree greater than or equal to k . Both the subspaces \mathcal{H}_k and the integers $\rho_k = \dim \mathcal{H}_k$ are invariant under regular static state feedback. There exists an integer $k^* > 0$ such that $\mathcal{H}_{k+1} = \mathcal{H}_k$ for $k \geq k^*$.

The following algorithm allows us to explicitly construct bases for the subspaces \mathcal{H}_k .

Step 1 Take $\{dx_1, \dots, dx_{l_1}, du_1, \dots, du_{l_2}\} \cup \{dx_1, \dots, dx_{l_1}\}$ as bases of \mathcal{H}_0 and \mathcal{H}_1 .

Step $k+1$ Suppose that $\{\eta_1, \dots, \eta_{l_{k-1}}, \mu_1, \dots, \mu_{l_k}\} \cup \{\eta_1, \dots, \eta_{l_{k-1}}\}$ are bases respectively of \mathcal{H}_{k-1} and \mathcal{H}_k and let us construct a basis for \mathcal{H}_{k+1} . The elements of \mathcal{H}_{k+1} are the one-forms $\omega \in \mathcal{H}_k$ such that $\omega \in \mathcal{H}_k$. Let $\omega = \sum_i \lambda_i \eta_i \in \mathcal{H}_k$ then $\omega = \sum_i (\lambda_i \eta_i + \lambda_j \eta_j)$. It is clear that $\omega \in \mathcal{H}_k$ if and only if $\sum_i \lambda_i \eta_i \in \mathcal{H}_k$. Now note that since $\eta_i \in \mathcal{H}_k$, η_j must be in \mathcal{H}_{k-1} so $\sum_i \lambda_i \eta_i$ may be written in the following form

$$\sum_i \lambda_i \eta_i = \sum_i \lambda_i \left(\sum_j \sigma_{ij} \eta_j + \sum_j \tau_{ij} \mu_j \right)$$

Thus $\omega \in \mathcal{H}_k$ if and only if the coefficients λ_j satisfy the following system of linear equations

$$\sum_j \tau_{ij} \lambda_j = 0 \quad 1 \leq i \leq \rho_{k-1} - \rho_k \quad (7)$$

This system of equations has $\rho_k - \rho$ linearly independent solutions being the rank of the matrix $[\tau_{ij}]$. Thus $\dim \mathcal{H}_{k+1}$ can be computed as

$$\alpha_k = \sum_{i=1}^{\rho_k} \lambda_i \eta_i \quad 1 \leq k \leq l$$

where $(\lambda_1', \dots, \lambda_{\rho_k}')^T$ are the ρ_{k+1} independent solutions of (7). The algorithm stops after a finite number k^* of steps when $\rho_{k+1} = 0$. In fact it is not difficult to show that $k^* \leq n - m + 1$.

Remark 3.2 In general the subspaces \mathcal{H}_k may depend on x, u in the following sense: the elements α built using the above algorithm can be written as linear combinations of dx, du and the coefficients being functions of $x, u, u^{(1)}, \dots$. However, a careful inspection of the construction shows that at the l th step these coefficients may be chosen to depend at most on u and its first $k-2$ time derivatives.

B. Accessibility Criteria

Proposition 3.3

1) \mathcal{H}_∞ is the largest subspace of $\mathcal{H}_l = \text{span}_{\mathcal{K}} \{dx, du\}$ which is invariant under time differentiation. It is also for any $k \geq 1$ the largest subspace of $\mathcal{H}_{l-1} = \text{span}_{\mathcal{K}} \{dx, du^{(1)}, \dots, du^{(k)}\}$ which is invariant under time differentiation.

2) Let $\{\alpha_1, \dots, \alpha_{l_\infty}\}$ be a basis for \mathcal{H}_∞ . Then the Frobenius condition $d\alpha_i \wedge \alpha_1 \wedge \dots \wedge \alpha_{l_\infty} = 0$, $1 \leq i \leq \rho_\infty$ is satisfied.

Proof Point 1 is a consequence of the construction (it is clear that starting from \mathcal{H}_{l-1} instead of \mathcal{H}_l in (6) one finds \mathcal{H}_∞ after k steps). The proof of point 2 is given in the Appendix. ■

In point 2 of Proposition 3.3 \wedge indicates the exterior product of differential forms. These conditions imply integrability of the Pfaffian system $\{\alpha_1, \dots, \alpha_{l_\infty}\}$ around regular points according to the dual version of Frobenius theorem. The reader who is not familiar with these matters is referred to [1].

Subspace \mathcal{H}_∞ may be interpreted as a codistribution on $\mathbb{R}^n \times \mathbb{R}^{m-1}$ where $k-1$ is the maximum number of input time derivatives necessary to write a basis of \mathcal{H}_∞ . Proposition 3.3 implies that this codistribution is locally integrable around any point where it has constant rank. This implies that \mathcal{H}_∞ is locally spanned by ρ_∞ exact one-forms $dx_1, \dots, dx_{l_\infty}$ where x_1, \dots, x_{l_∞} are functions defined around such regular points. These functions do not depend on u and its time derivatives because $\mathcal{H}_\infty \subset \text{span}_{\mathcal{K}} \{dx\}$ so \mathcal{H}_∞ may be interpreted as a codistribution on \mathbb{R}^n . Since for a function $v(x)$, $dv \in \mathcal{H}_\infty$ is equivalent to v being constant along all the vector fields $\text{ad}_j^{l_k}(f)$ with $j \geq 0$ and g_k the control vector fields (i.e. to be constant along the strong accessibility distribution) \mathcal{H}_∞ is around regular points the annihilator of the strong accessibility distribution. This leads to the following

Proposition 3.4 (Accessibility Criteria) The following statements are equivalent

- 1) System Σ satisfies the strong accessibility rank condition
- 2) Any nonzero one-form has finite relative degree
- 3) $\mathcal{H}_\infty = \{0\}$

Proposition 3.4 is one key result of this note since it allows the following construction which we then relate to linearizing outputs if these exist.

C. The Infinitesimal Brunovsky Form

Theorem 3.5 Suppose $\mathcal{H}_\infty = \{0\}$. There exists a list of integers $\{r_1, \dots, r_l\}$ invariant under regular static state feedback and m one-forms $\omega_1, \dots, \omega_m$ with relative degrees r_1, \dots, r_l such that

- 1) $\text{span}_\Lambda \{\omega_i^{(j)}, 1 \leq i \leq m, 0 \leq j \leq r_i - 1\} = \text{span}_\Lambda \{dx\}$;
 - 2) $\text{span}_\Lambda \{\omega_i^{(j)}, 1 \leq i \leq m, 0 \leq j \leq r_i\} = \text{span}_\Lambda \{dx, du\}$;
 - 3) the forms $\{\omega_i^{(j)}, 1 \leq i \leq m, j \geq 0\}$ are linearly independent.
- In particular, $\sum_{i=1}^m r_i = n$.

Proof: Let \mathcal{W}_k be a basis for \mathcal{H}_k . By definition, \mathcal{W}_k and $\dot{\mathcal{W}}_k$ are in \mathcal{H}_{k+1} . Note that \mathcal{W}_k and $\dot{\mathcal{W}}_k$ are linearly independent. For, let $\mathcal{W}_k = \{\eta_1, \dots, \eta_{p_k}\}$ —then $\dot{\mathcal{W}}_k = \{\dot{\eta}_1, \dots, \dot{\eta}_{p_k}\}$ —and suppose that there exist some coefficients $\{\lambda_i, \mu_i\}$ such that $\sum_i (\lambda_i \eta_i + \mu_i \dot{\eta}_i) = 0$. The linear independence of the η_i 's implies that not all the μ_i 's vanish. Now consider the one-form $\omega = \sum_i \mu_i \eta_i$ whose time derivative is $\dot{\omega} = \sum_i \dot{\mu}_i \eta_i - \sum_i \lambda_i \eta_i$. This implies that ω is in \mathcal{H}_{k+1} , which is a contradiction. Hence, it is always possible to choose a set \mathcal{W}_{k+1} (possibly empty) such that $\mathcal{W}_k, \dot{\mathcal{W}}_k, \mathcal{W}_{k+1}$ is a basis for \mathcal{H}_{k+1} . This procedure is repeated k^* times, so the sequence $\{\mathcal{H}_k\}$ is shown to have the following structure:

$$\mathcal{H}_k = \text{span}_\Lambda \{\omega_i^{(j)}, k \leq i \leq k^*, 0 \leq j \leq i-k, 0 \leq k \leq k^*.$$

$\mathcal{H}_1 = \text{span}_\Lambda \{dx\}$ and $\text{rank } g(x) = m$ imply that $\mathcal{W}_0 = \emptyset$. It can be proved by induction that, for $0 \leq k \leq k^*$, the set $\{\mathcal{W}_k, \dots, \mathcal{W}_{k^*}^{(k^*-k)}, \dots, \mathcal{W}_k\}$ is linearly independent. Finally, set $\{\omega_1, \dots, \omega_m\} = \{\mathcal{W}_k, \dots, \mathcal{W}_1\}$. The invariance of the list of relative degrees is rather obvious from the construction. It can also be seen from the fact that the number of r_i 's which are equal to k is given by $\kappa_k = \text{card } \mathcal{W}_k = \dim \mathcal{H}_k / (\mathcal{H}_{k+1} + \dot{\mathcal{H}}_{k+1})$. ■

The following (straightforward) corollary of Theorem 3.5 is the reason for the name infinitesimal Brunovsky form: the ω_k 's provide a basis of one-forms in which the first-order approximation of Σ looks like a linear Brunovsky canonical form [11]. If these forms were integrable, then they would yield a true Brunovsky canonical form for system Σ (for this reason, the term infinitesimal is preferred to "non-exact," used in [28]). In any case, the integers r_i are nice candidates for generalizing to nonlinear systems the notion of linear Kronecker controllability indices.

Corollary 3.6 (The Infinitesimal Brunovsky form): Suppose $\mathcal{H}_\infty = \{0\}$. Then the basis $\{\omega_i^{(j)}, 1 \leq i \leq m, 1 \leq j \leq r_i\}$ of $\text{span}_\Lambda \{dx\}$ defined by $\omega_i^{(j)} = \omega_i^{(j-1)}$ yields

$$\begin{aligned} \dot{\omega}_{i-1} &= \omega_{i-2} \\ &\vdots \\ \dot{\omega}_{i-1} &= \omega_{i-1} \\ \dot{\omega}_{i-1} &= \sum_{j=1}^n a_{i,j} dx_j + \sum_{j=1}^m b_{i,j} du_j \end{aligned} \quad (1 \leq i \leq m)$$

where $a_{i,j}, b_{i,j} \in \mathcal{K}$ and $[b_{i,j}]$ has an inverse in the ring of $m \times m$ matrices with entries in \mathcal{K} .

IV. MAIN RESULTS

A. Some Preliminaries

An Algebra of Polynomial Operators: Let $\mathcal{K}[d/dt]$ denote the (noncommutative) algebra of polynomials in the operator d/dt with coefficients in \mathcal{K} . The addition and external multiplication are the usual ones. The internal multiplication corresponds to operators composition: $(d/dt)(p) = p(d/dt) + \dot{p}$, $\forall p \in \mathcal{K}$. The only invertible elements in $\mathcal{K}[d/dt]$ are the nonzero elements of \mathcal{K} (i.e., nonzero polynomials of degree zero). Let $\mathcal{K}^{m \times m}[d/dt]$ denote the algebra of $m \times m$ matrices with entries in $\mathcal{K}[d/dt]$. Let \mathcal{E}^m be the differential \mathcal{K} -vector space spanned by m -tuples of one-forms. Each $P \in \mathcal{K}^{m \times m}[d/dt]$ defines a differential operator in \mathcal{E}^m : $P \cdot \Omega = \sum_i P_i \Omega_i^{(i)}$ for all $\Omega = (\omega_1, \dots, \omega_m)^T \in \mathcal{E}^m$, where $\Omega_i^{(i)} = (\omega_1^{(i)}, \dots, \omega_m^{(i)})^T$ and $P_i := \sum_j P_{ij}(d/dt)^j \in \mathcal{K}^{m \times m}[d/dt]$.

Invertible elements of $\mathcal{K}^{m \times m}[d/dt]$ play an important role; they are elements $P \in \mathcal{K}^{m \times m}[d/dt]$ such that there exists $Q \in \mathcal{K}^{m \times m}[d/dt]$ such that $P \cdot Q = Q \cdot P = I_m$.

Definition 4.1 (Structure at Infinity for One-Forms): $\Omega = (\omega_1, \dots, \omega_m)^T$ is said to have $\sigma_k = \dim(\text{span}_\Lambda \{dx, \Omega, \dots, \Omega^{(k)}\} / \text{span}_\Lambda \{dx, \Omega, \dots, \Omega^{(k-1)}\})$ zeros at infinity of order less than or equal to k .

For exact one-forms ($\omega_i = dh_i$), Definition 4.1 coincides with (3) and (4), since exterior differentiation and time-differentiation commute, so that the notations of Section II may be adopted verbatim. In particular, Lemma 2.1 is also valid for systems of one-forms.

Proposition 4.2: Consider the system of m one-forms $\Omega := (\omega_1, \dots, \omega_m)^T$, and the polynomial matrix operator $P \in \mathcal{K}^{m \times m}[d/dt]$. Let $\tilde{\Omega} := P \cdot \Omega$. Then

$$\dim(\mathcal{N}_\nu \cap \text{span}_\Lambda \{\Omega^{(k)}, k \geq 0\}) \leq \dim(\mathcal{N}_\nu \cap \text{span}_\Lambda \{\tilde{\Omega}^{(k)}, k \geq 0\})$$

where ν is an integer large enough such that Ω and $\tilde{\Omega}$ belong to \mathcal{N}_ν .

Proof: Suppose P has degree α . Straightforward computations show that $\tilde{\Omega}^{(k)}$, for $k \geq 0$, can be written as a linear combination of the following form $\tilde{\Omega}^{(k)} = R_0 \Omega + R_1 \dot{\Omega} + \dots + R_{k+\alpha} \Omega^{(k+\alpha)}$. Thus, $\text{span}_\Lambda \{\tilde{\Omega}^{(k)}, k \geq 0\} \subseteq \text{span}_\Lambda \{\Omega^{(k)}, k \geq 0\}$ and the result follows. ■

B. The Results

Our main result is the following. It is an easy consequence of Theorem 3.5 and Proposition 4.2.

Theorem 4.3 (Problem Statement 2): Suppose $\mathcal{H}_\infty = \{0\}$. There exists a system of linearizing outputs if and only if there exists an invertible polynomial operator $P \in \mathcal{K}^{m \times m}[d/dt]$ such that $d(P\Omega) = 0$, where $\Omega = (\omega_1, \dots, \omega_m)^T$ is a system of one-forms characterized by Theorem 3.5.

Proof:

Necessity: Suppose $y = h(x, u, \dots, u^{(\nu-1)})$ is a linearizing output. Problem Statement 2 implies that $\mathcal{E} = \mathcal{Y}$. Theorem 3.5 implies that $\mathcal{E} = \text{span}_\Lambda \{\Omega^{(k)}, k \geq 0\}$. Thus, there exist polynomial matrix operators P, Q such that $dy = P\Omega$ and $\Omega = Q dy$. Clearly, $PQ = QP = I_m$ and hence P is invertible. Moreover, $d(P\Omega) = d(dy) = 0$.

Sufficiency: Let $N = \dim(x_\nu \cap \text{span}_\Lambda \{\Omega^{(k)}, k \geq 0\})$, $\tilde{N} = \dim(\mathcal{N}_\nu \cap \text{span}_\Lambda \{\Omega^{(k)}, k \geq 0\})$, where $\Omega = P\Omega$. Theorem 3.5 implies that $N = n + m\nu$. Existence of the operator P implies $N \leq \tilde{N}$. Invertibility of P implies the existence of an operator Q such that $\Omega = Q\tilde{\Omega}$, i.e., $N \leq \tilde{N}$ and hence $N = \tilde{N}$. The result follows because one can assume, without loss of generality, that $\Omega = d(x, u, \dots, u^{(\nu-1)})$, y is a linearizing output. ■

Theorem 4.3 relates linearizing outputs, if they exist, to the set of differential one-forms built in Theorem 3.5 for arbitrary accessible systems. It provides an alternative way to tackle the problem of deciding whether linearizing outputs exist, i.e., whether a given system is linearizable by endogeneous dynamic feedback, by looking for an invertible matrix P meeting the above conditions. This does not provide a practically checkable criterion because the degree (in the operator d/dt) of the matrix P is not known *a priori*, which prevents the condition of the theorem from being finitely checkable. By forcing P to have degree zero (i.e., to be an invertible matrix with entries in \mathcal{K}), the problem is made finite, and one obtains the following sufficient condition.

Corollary 4.4: A sufficient condition for the existence of a system of linearizing outputs is that a system of one-forms $\Omega = (\omega_1, \dots, \omega_m)^T$ satisfying the conditions of Theorem 3.5 satisfy the Frobenius condition

$$d\omega_i \wedge \omega_1 \wedge \dots \wedge \omega_m = 0, \quad 1 \leq i \leq m. \quad (8)$$

The relative degrees of $\omega_1, \dots, \omega_m$ coincide with the orders of the zeros at infinity of the linearizing outputs.

Proof: The Frobenius condition implies that there exists a basis composed of exact one-forms for the codistribution spanned by $\{\omega_1, \dots, \omega_m\}$, and hence that there exists an invertible matrix (with entries in \mathcal{K}) relating this basis to $\{\omega_1, \dots, \omega_m\}$. ■

The condition (8) is obviously finitely (and easily) checkable once some ω_i 's have been constructed. It is of course not a necessary condition, and it should be noted that, in general, it depends on the choice of the ω_i 's: some systems of one-forms satisfying the conditions of Theorem 3.5 may satisfy the Frobenius condition (8), whereas some others do not. Even for a linear system, a wrong choice of the one-forms $\omega_1, \dots, \omega_m$ prevents condition (8) from being satisfied. However, in many practical cases (see the proofs of the results in Section V or [28, Section 3.2]), it is not difficult to round this difficulty and check whether (8) is met for one of the possible choices of the ω_i 's.

A way for bounding the degree of P is to look for linearizing outputs depending on x only, as illustrated by the following result.

Theorem 4.5 (Problem Statement 1): Suppose $\mathcal{H}_\infty = \{0\}$. Then there exists a system of linearizing outputs which depend only on x if and only if the conditions of Theorem 4.3 are satisfied and, in addition, $\deg(P_{ij}(d/dt)) \leq r_j - 1$, $1 \leq i, j \leq m$.

Proof: Sufficiency is obvious. Conversely, suppose that one of the polynomial elements of the matrix $P(d/dt)$, say $P_{ij}(d/dt)$ has degree equal to r_j . Thus, dh_i contains a term which depends on du and that cannot be eliminated by the remaining terms since, by construction, all the $\omega_i^{(k)}$ are linearly independent. This is a contradiction. ■

Note that it is very easy to write down some similar criteria for the existence of linearizing outputs depending on x , u , and any finite number of time derivatives of u . Such types of conditions as the ones given in Theorem 4.5 can be restated as existence of a finite number of functions—the coefficients of the polynomial entries of P —meeting some differential conditions, namely, $d(P\Omega) = 0$ and P invertible. A possible way to avoid writing the relations on the entries of P for it to be invertible is to write P as a finite product of elementary invertible matrices and taking the coefficients of these elementary matrices as unknowns instead of the entries of P itself; this is exploited in [27]. Checking whether there exist some linearizing outputs depending on x , u , and any finite number of time derivatives of u therefore amounts to checking whether a finite set of PDE's in a finite number of unknown functions has a solution.

This is not new since it is easy (although tedious) to write down the PDE's which have to be satisfied by the linearizing outputs themselves, if they are restricted to depend on x only. This is the underlying idea of the characterizations given in some particular cases (see, for instance, [25]). We, however, believe that looking for the invertible matrix P once the ω_k 's have been constructed is more natural and more tractable. This is illustrated by the very short proofs of the theorems of next section, which are known but usually not so natural to prove, and by results like the ones obtained in [28, Section 3.2], [27], [3, Theorem 5.4], which work out some nontrivial particular cases.

V. PARTICULAR CASES

In this section we recover some classical results using the infinitesimal Brunovsky form.

Theorem 5.1 (Static State Feedback Linearization): System Σ is linearizable by static state feedback if, and only if, $\mathcal{H}_\infty = \{0\}$ and, for $k = 1, \dots, k^*$, \mathcal{H}_k is completely integrable.

This is, of course, equivalent to the early characterization [17] and [22], or to the more recent one given in [17] and [22]. The infinitesimal Brunovsky form provides a very short proof.

Proof: If each \mathcal{H}_k is completely integrable, one may choose forms, say, $\omega_i = dq_i(x)$, $i = 1, \dots, m$ in Theorem 3.5 whose m functions $q_i(x)$ whose relative degrees satisfy $\sum_{i=1}^m r_i = n$ whose decoupling matrix [21] is nonsingular because the ω_i 's and their time derivatives are linearly independent, this completes the first part. The converse is obvious since the \mathcal{H}_k 's are invariant by static feedback, and are integrable for a linear system. ■

Theorem 5.2 (Single-Input Systems). Let Σ be a single-input system and suppose $\mathcal{H}_\infty = \{0\}$. Then there is only one differential form ω_1 in Theorem 3.5, and the following statements are equivalent:

1) Σ is linearizable by static state feedback; 2) Σ is linearizable by dynamic state feedback; 3) $d\omega_1 \wedge \omega_1 = 0$, where ω_1 is such that $\mathcal{H}_0 = \text{span}_{\mathcal{K}} \{\omega_1\}$.

This result (equivalence between 1 and 2) was first obtained in [6] and [7]. The infinitesimal Brunovsky form—note that ω_1 is invariant up to a nonzero multiplicative function—allows us to give the simple characterization 3 and the following very simple proof.

Proof: It is obvious from Theorem 5.1 and Theorem 4.3 because the only invertible elements of $\mathcal{K}[d/dt]$ are those having degree 0. In turn, multiplication by a nonzero function does not change the rank of the differential form $d\omega_1 \wedge \omega_1$.

Theorem 5.3 (Systems with $m = n - 1$ inputs [7]): A system Σ with $m = n - 1$ inputs is linearizable by dynamic state feedback if, and only if, $\mathcal{H}_\infty = \{0\}$.

Proof: \mathcal{H}_2 is generated by a single nonzero one-form ω_1 which is orthogonal to the distribution spanned by the vector fields g_1, \dots, g_{n-1} , and thus can be chosen independent of u . $\omega_2, \dots, \omega_{n-1}$ can be chosen arbitrarily, linearly independent of $\{\omega_1, \omega_2\}$ and belonging to $\text{span}_{\mathcal{K}} \{dx\}$: they can also be chosen independent of u . Then, the differential forms $d\omega_i \wedge \omega_1 \wedge \dots \wedge \omega_{m-1}$, $i = 1, \dots, m$, are zero because they are $(n+1)$ -forms in n variables. The converse is obvious. ■

VI. CONCLUSION

We have built a so-called infinitesimal Brunovsky form which exhibits m controllable blocks whose dimensions play the role of Kronecker controllability indices in the linear case. This extension to the nonlinear case is innovative since, to our best knowledge, it is the only available one with the property that the sum of these indices equals the dimension of the strong accessibility distribution. This result on nonlinear accessibility was used to derive a necessary and sufficient condition for existence of a linearizing output, and the early results in (either static or dynamic) feedback linearization have been shown to fit naturally in our formalism. Static feedback linearization was shown to be a matter of exact one-forms whereas dynamic feedback linearization is a matter of possibly nonexact one-forms.

APPENDIX

PROOF OF PROPOSITION 3.3, POINT 2

Define \mathcal{E}^* to be the dual vector space of \mathcal{E} (topology induced by $\|\sum a_j dv_j\|^2 = \sum a_j^2$, dv_j 's taken among the dx_i 's or the $du_k^{(j)}$'s, both sums are finite), whose elements—"vector fields"—are of the form

$$X = \sum_{i=1}^n a_i \frac{\partial}{\partial x_i} + \sum_{k=1}^m \sum_{j \geq 0} b_{j,k} \frac{\partial}{\partial u_k^{(j)}}$$

where $a_i, b_{j,k}$ are in \mathcal{K} and $\partial/\partial x_i, \partial/\partial u_k^{(j)}$ are defined by $\langle \partial/\partial x_i, dx_i \rangle = \delta_{i,i'}$, $\langle \partial/\partial u_k^{(j)}, du_k^{(j')} \rangle = \delta_{j,j'} \delta_{k,k'}$, $\langle \partial/\partial u_k^{(j)}, \dots \rangle$

$d\epsilon_i = \langle \partial/\partial x_i, du_k^{(j)} \rangle = 0$. Note that the above linear combination need not be finite, but does define an element of \mathcal{E}^* because, for all $\omega \in \mathcal{E}$, $\langle X, \omega \rangle$ may be written as a sum with only finitely many nonzero terms. Define the "time-derivative" of $X \in \mathcal{E}$ by

$$\langle \dot{X}, \omega \rangle = \langle \dot{X}, \omega \rangle - \langle X, \dot{\omega} \rangle. \quad (9)$$

The interior product (or hook) $i(X)$ is defined, for a differential form of degree two $\mu = \sum_{i,j} a_{i,j} dv_i \wedge dv_j$, by $i(X)\mu = \sum_{i,j} a_{i,j} (\langle X, dv_i \rangle dv_j - \langle X, dv_j \rangle dv_i)$. Clearly, taking the time-derivative of both sides in this identity yields, from (9),

$$\dot{i(X)\mu} = i(\dot{X})\mu + i(X)\dot{\mu}. \quad (10)$$

Let $\alpha_1, \dots, \alpha_{p_\infty}$ be a basis of \mathcal{H}_∞ , and define the following subspace of \mathcal{E}^*

$$\mathcal{G}_\infty = \{X \in \mathcal{E}^* \mid \forall \omega \in \mathcal{H}_\infty, \langle X, \omega \rangle = 0 \text{ and } \alpha_1 \wedge \dots \wedge \alpha_{p_\infty} \wedge i(X)d\omega = 0\}$$

whose elements are sometimes called [4] the "Cauchy characteristic vector fields" of \mathcal{H}_∞ . The "characteristic system" (or "retracting space" according to [4]) of \mathcal{H}_∞ is defined as its annihilator: $\mathcal{C}(\mathcal{H}_\infty) = \mathcal{G}_\infty^\perp$. For a certain $K \geq 0$, the elements $\alpha_1, \dots, \alpha_{p_\infty}$ of the chosen basis for \mathcal{H}_∞ can be written as linear combinations of dx_1, \dots, dx_n with coefficients function of $x, u, \dot{u}, \dots, u^{(K)}$ only, hence all the α_i 's are linear combinations of elements of the form $dx_j \wedge dx_l$ or $dx_j \wedge du_k^{(l)}$ with $l \leq K$, so that all the elements $\partial/\partial u_k^{(j)}$ are in \mathcal{G}_∞ for $j \geq K+1$, and finally that $\mathcal{C}(\mathcal{H}_\infty)$ is a subspace of $\text{span}_K \{dx, du, \dot{u}, \dots, u^{(K)}\}$. Furthermore, we have the following.

Lemma: $\mathcal{C}(\mathcal{H}_\infty)$ is invariant by time-differentiation.

From point 1 of Proposition 3.3, this implies $\mathcal{C}(\mathcal{H}_\infty) \subset \mathcal{H}_\infty$, and hence, from standard exterior algebra, $\alpha_1 \wedge \dots \wedge \alpha_{p_\infty} \wedge i(X)d\omega = 0$ for $i = 1, \dots, p_\infty$.

Proof of the Lemma: Consider $X \in \mathcal{G}_\infty$, $\omega \in \mathcal{H}_\infty$. We have [compare (9) and (10)]

$$\langle \dot{X}, \omega \rangle = \langle \dot{X}, \omega \rangle - \langle X, \dot{\omega} \rangle$$

$$\begin{aligned} & \alpha_1 \wedge \dots \wedge \alpha_{p_\infty} \wedge i(\dot{X})d\omega \\ &= \overline{\alpha_1 \wedge \dots \wedge \alpha_{p_\infty} \wedge i(X)d\omega} - \alpha_1 \wedge \dots \wedge \alpha_{p_\infty} \wedge i(X)d\omega - \dots \\ & - \alpha_1 \wedge \dots \wedge \alpha_{p_\infty} \wedge i(X)d\omega - \alpha_1 \wedge \dots \wedge \alpha_{p_\infty} \wedge i(X)d\omega. \end{aligned}$$

Since $\langle X, \omega \rangle$ and $\alpha_1 \wedge \dots \wedge \alpha_{p_\infty} \wedge i(X)d\omega$ are identically zero, and $\dot{\alpha}_1, \dots, \dot{\alpha}_{p_\infty}$ and $\dot{\omega}$ are in \mathcal{H}_∞ (because \mathcal{H}_∞ is invariant by time-differentiation), all the terms on the right-hand sides above are zero, which implies $\langle \dot{X}, \omega \rangle = 0$ and $\alpha_1 \wedge \dots \wedge \alpha_{p_\infty} \wedge i(\dot{X})d\omega = 0$. Hence, \dot{X} is in \mathcal{G}_∞ because this is true for all $\omega \in \mathcal{H}_\infty$. Now, take η

in $\mathcal{C}(\mathcal{H}_\infty)$. For all $X \in \mathcal{G}_\infty$, $\langle X, \eta \rangle = \langle \dot{X}, \eta \rangle - \langle X, \dot{\eta} \rangle = 0$ where $\langle \dot{X}, \eta \rangle$ is zero because $\langle X, \eta \rangle$ is identically zero, and $\langle \dot{X}, \eta \rangle$ is zero because \dot{X} is in \mathcal{G}_∞ . This proves $\eta \in \mathcal{C}(\mathcal{H}_\infty) \Rightarrow \dot{\eta} \in \mathcal{C}(\mathcal{H}_\infty)$, and therefore the lemma. ■

REFERENCES

- [1] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis and Applications*, 2nd. ed., Applied Mathematical Sciences, vol. 75, New York: Springer, 1988.
- [2] E. Aranda-Bricaire, C. H. Moog, and J.-B. Pomet, "Infinitesimal Brunovsky form for nonlinear systems with applications to dynamic

linearization," presented at the Ext. Workshop Geometry Nonlinear Control, Warsaw, June 1993; *Proceedings*, Banach Center Publications, to appear.

- [3] —, "Feedback linearization: A linear algebraic approach," in *Proc. 32nd IEEE Conf. Decision Contr.*, 1993, pp. 3441–3446.
- [4] R. L. Bryant, S. S. Chern, R. B. Gardner, H. L. Goldschmitt, and P. A. Griffiths, *Exterior Differential Systems*, Mathematical Sciences Research Institute Publications, vol. 18, New York: Springer, 1991.
- [5] L. Cao and Y. F. Zheng, "On minimal compensators for decoupling control," *Syst. Contr. Lett.*, vol. 18, pp. 121–128, 1992.
- [6] B. Charlet, J. Lévine, and R. Marino, "On dynamic feedback linearization," *Syst. Contr. Lett.*, vol. 13, pp. 143–151, 1989.
- [7] —, "Sufficient conditions for dynamic feedback linearization," *SIAM J. Contr. Opt.*, vol. 29, pp. 38–57, 1991.
- [8] D. Claude, "Everything you always wanted to know about linearization," in *Algebraic and Geometric Methods in Nonlinear Control Theory*, M. Fliess and M. Hazewinkel, Eds. Dordrecht, The Netherlands: Reidel, 1986, pp. 181–220.
- [9] M. D. Di Benedetto, J. Grizzle, and C. H. Moog, "Rank invariants of nonlinear systems," *SIAM J. Contr. Opt.*, vol. 27, pp. 658–672, 1989.
- [10] M. Fliess, "A new approach to the noninteracting control problem in nonlinear system theory," in *Proc. 23rd Allerton Conf.*, Monticello, 1985, pp. 123–129.
- [11] M. Fliess, "Some remarks on the Brunovsky canonical form," *Kybernetika*, vol. 29, pp. 417–422, 1993.
- [12] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, "On differentially flat nonlinear systems," in *Proc. 2nd IFAC NOLCOS Symp.*, M. Fliess, Ed., Bordeaux, 1992, pp. 408–412.
- [13] —, "Sur les systèmes non linéaires différentiellement plats," *C. R. Acad. Sci. Paris*, vol. 315-1, pp. 619–624, 1992.
- [14] —, "Linéarisation par bouclage dynamique et transformations de Lie-Bäcklund," *C. R. Acad. Sci. Paris*, vol. 317-1, pp. 981–986, 1993.
- [15] —, "Towards a new differential geometric setting in nonlinear control," presented at the *Int. Geometrical Colloq.*, Moscow, May 1993; to appear in the *Proceedings*.
- [16] —, "Flatness and defect of nonlinear systems: Introductory theory and examples," *Int. J. Contr.*, to appear.
- [17] R. B. Gardner and W. F. Shadwick, "The GS algorithm for exact linearization to Brunovsky normal form," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 224–230, 1992.
- [18] L. R. Hunt, R. Su, and G. Meyer, "Design for multi input nonlinear systems," in *Differential Geometric Control Theory*, R. Brockett, R. Millmann, and H. Sussmann, Eds. Boston: Birkhäuser, 1983, pp. 268–298.
- [19] A. Isidori, C. H. Moog, and A. De Luca, "A sufficient condition for full linearization via dynamic state feedback," in *Proc. 25th IEEE Conf. Dec. Contr.*, 1986, pp. 203–208.
- [20] A. Isidori and C. H. Moog, "On the nonlinear equivalent of the notion of transmission zeros," in *Modeling and Adaptive Control*, Lecture Notes in Control and Information Science, vol. 105, C. I. Byrnes, and A. Kurzhanski, Eds. New York: Springer-Verlag, 1988, pp. 146–158.
- [21] A. Isidori, *Nonlinear Control Systems*, 2nd ed., Communications and Control Engineering Series. New York: Springer-Verlag, 1989.
- [22] B. Jancubczyk and W. Respondek, "On linearization of control systems," *Bull. Acad. Pol. Sci., Ser. Sci. Math.*, vol. 18, pp. 517–522, 1980.
- [23] B. Jancubczyk, "Remarks on equivalence and linearization of nonlinear systems," in *Proc. 2nd IFAC NOLCOS Symp.*, M. Fliess, Ed., Bordeaux, 1992, pp. 393–397.
- [24] P. Martin, "Contribution à l'étude des systèmes non linéaires différentiellement plats," Thèse de Doctorat, Ecole des Mines de Paris, 1992.
- [25] —, "A geometric sufficient condition for flatness of systems with m inputs and $m+1$ states," in *Proc. 32 IEEE Conf. Dec. Contr.*, pp. 3431–3436, 1993.
- [26] C. H. Moog, "Nonlinear decoupling and structure at infinity," *Math. Contr. Signals Syst.*, vol. 1, pp. 257–268, 1988.
- [27] J.-B. Pomet, "On dynamic feedback linearization of four-dimensional affine control systems with two inputs," preprint, 1993.
- [28] J. B. Pomet, C. H. Moog, and E. Aranda, "A non-exact Brunovsky form and dynamic feedback linearization," in *Proc. 31st IEEE Conf. Dec. Contr.*, pp. 2012–2017, 1992.
- [29] J.-B. Pomet, "A differential geometric setting for dynamic equivalence and dynamic linearization," presented at the Ext. Workshop Geometry Nonlinear Control, Warsaw, June 1993; *Proceedings*, Banach Center Publications, to appear.
- [30] W. M. Sluis, "Absolute equivalence and its applications to control theory," Ph.D. dissertation, Univ. Waterloo, 1992.

Design of a Class of Luenberger Observers for Descriptor Systems

M. Hou and P. C. Müller

Abstract—A new approach to the design of a class of Luenberger observers for descriptor systems is presented. The design is performed in a straightforward manner. The conditions required in the design are given in terms of the original system matrices. The relationship among these and other well-known conditions is highlighted.

I. INTRODUCTION

The design of Luenberger-type observers for descriptor systems has obtained considerable attention in the past decade, see, e.g., [1]–[11].

Descriptor systems under consideration have the following form:

$$E\dot{x} = Ax + Bu \quad (1)$$

$$y = Cx + Du \quad (2)$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^r$, $y \in \mathbb{R}^m$ are the descriptor variables, control, and measurement vectors, respectively. E , $A \in \mathbb{R}^{n \times n}$, B , C , and D are known real matrices of appropriate dimensions. $\text{rank } E = r$, $0 < r < n$.

Mostly, the matrices E and A are assumed to be square, i.e., $q = n$, and the matrix pencil $\lambda E - A$ is required to be regular. However, as shown in [3] and [10], concerning observer design this kind of assumption is not needed. Therefore, no assumption on $\lambda E - A$ is made in this note.

We accept piecewise continuous distributions as solutions [12]. It is well known that arbitrary initial conditions $E x(0_-)$ and inputs $u(t)$ are admissible for a descriptor system with a regular matrix pencil. However, since no assumption on $\lambda E - A$ has been made in (1), we should assume that u and a certain but unknown initial condition $E x(0_-)$ are admissible, i.e., they do not result in contrary equations in (1). More precisely, $E x(0_-)$ and $u(t)$ are said to be admissible if they satisfy $\text{rank} [-sE + A] E x(0_-) B U(s) = \text{rank} (sE - A)$ for some $s \in \mathbb{C}$, where $U(s)$ is the Laplace transformation of $u(t)$. Thus, there exists at least one trajectory obeying (1) and (2). It is worth pointing out that the admissibility and consistency of the initial condition $E x(0_-)$ are two different concepts. The former implies the existence of solution(s), while the latter guarantees further the impulse-free behavior of the solution(s).

Certainly, if $\lambda E - A$ is not of full column rank, i.e., for any $\lambda \in \mathbb{C}$ $\text{rank}(\lambda E - A) < n$, then there exist more than one trajectory satisfying (1). Here, we emphasize the difference between the description of a system, e.g., (1), and the real system. Although the description (1) is of course assumed to be correct, it may be incomplete. The column singularity of $\lambda E - A$ reflects incompleteness of the description. In such a case, one cannot answer the question of uniqueness of trajectories only from (1). However, if

$$\begin{bmatrix} -\lambda E + A \\ C \end{bmatrix}$$

is column regular, i.e., there exists $\lambda \in \mathbb{C}$ such that this extended matrix pencil has rank n , then the solution $x(t)$ satisfying (1)

Manuscript received August 26, 1993; revised November 18, 1993, March 14, 1994, and May 29, 1994.

The authors are with Sicherheitstechnische Regelungs- und Meßtechnik, Bergische Universität-GH Wuppertal, D-42097 Wuppertal, Germany.

IEEE Log Number 9407026.

and (2) for any given admissible initial condition are unique. By treating y formally as some passive inputs, this can be proven following the analysis by, for example, Gantmacher and Flether [14].

In the observer design for descriptor systems, the condition

$$\text{rank} \begin{bmatrix} -\lambda E + A \\ C \end{bmatrix} = n$$

for any $\lambda \in \mathbb{C}$, $\text{Re}(\lambda) > 0$ is needed at any rate. This condition implies column regularity of

$$\begin{bmatrix} -\lambda E + A \\ C \end{bmatrix}$$

and therefore uniqueness of the trajectory. This means that once an observer for a descriptor system is obtained, the observer supplies the estimation of a trajectory, and this trajectory of the underlying system has to be unique even if uniqueness of the trajectory may not be recognized from (1) alone before.

Efforts have been made to design observers for descriptor systems under different conditions using various techniques. A series of results of Paraskevopoulos *et al.* [7], [8], [15] devote to the design of a class of Luenberger observers of the form

$$\dot{z} = E_1 z + E_2 y + E_3 u, \quad z \in \mathbb{R}^p \quad (3)$$

$$x = F_1 z + F_2 y \quad (4)$$

where $\hat{x}(t)$ goes to $x(t)$ as t tends to infinity for any initial condition $z(0)$ and arbitrary admissible $E x(0_-)$. The underlying descriptor systems in [7] and [8] are of the form (1), (2) with the assumptions $|\lambda E - A| \neq 0$ and $D = 0$.

The more general Luenberger observer has the same form as (3) and (4) except that (4) also contains the input term $F_3 u$. Generally, this kind of observer needs less restrictive conditions than that of the above class of Luenberger observers as shown in [16]. The research is progressing toward achievement of necessary and sufficient existence conditions as well as design procedures of Luenberger observers for descriptor systems. Although for the class of Luenberger observers of the form (3), (4) an effort has been made in [7], [8] to approach this goal, the resulted existence conditions can be proved to be sufficient only [10], [16]. On the other hand, the design in [7], [8] is rather involved.

The aim of this note is to propose a straightforward design of the observers described by (3) and (4), and to provide the corresponding conditions in terms of the original system matrices. The treatment in this note permits the general descriptor systems like (1), (2) to be taken into consideration. The existence conditions in the simple form enable us to compare them to the existing ones.

Finally, it is worth noting that a natural choice of observer form may be

$$E\dot{\hat{x}} = A\hat{x} + Bu + L(y - C\hat{x} - Du). \quad (5)$$

Thus, the design of Luenberger observers can be addressed as finding L such that in the error equation $E\dot{e} = (A + LC)e$, where $e = x - \hat{x}$, e is free of impulsive behavior and goes to zero as t tends to infinity. This goal is nevertheless not reachable if E and A have fewer rows than columns. Therefore, the form (5) is not recommended for dealing with the observer design for descriptor systems of the type (1), (2).

II. EXISTENCE CONDITIONS AND DESIGN PROCEDURE

The main result of this note is as follows. Its proof provides a design procedure.

Theorem 2.1: There exists an observer of the type (3), (4) if the following two conditions are fulfilled:

$$\text{rank} \begin{bmatrix} -\lambda E + A \\ C \end{bmatrix} = n, \quad \forall \lambda \in \mathbb{C}, \quad \text{Re}(\lambda) \geq 0, \quad (6)$$

$$\text{rank} \begin{bmatrix} E & A & B \\ 0 & C & D \\ 0 & E & 0 \end{bmatrix} = n + \text{rank} \begin{bmatrix} E & B \\ 0 & D \end{bmatrix}. \quad (7)$$

Proof: Using the singular value decomposition technique [18], one can always find two nonsingular orthogonal matrices $\tilde{P} \in \mathbb{R}^{n \times n}$ and $\tilde{Q} \in \mathbb{R}^{n \times n}$ such that $\tilde{P}E\tilde{Q} = \text{diag}(\Sigma, 0)$, where Σ , is nonsingular. Thus, by choosing $P = \text{diag}(\Sigma^{-1}, I_{q-m})\tilde{P}$, one has

$$\begin{aligned} \text{diag}(P, I_m) \begin{bmatrix} E & A & B \\ 0 & C & D \end{bmatrix} \text{diag}(\tilde{Q}, \tilde{Q}, I_s) \\ = \begin{bmatrix} I_r & 0 & A_{11} & A_{12} & B_{11} \\ 0 & 0 & A_{21} & A_{22} & B_{21} \\ 0 & 0 & C_{11} & C_{22} & D \end{bmatrix}. \end{aligned} \quad (8)$$

We can further find an orthogonal matrix $P_1 \in \mathbb{R}^{(q-m) \times (q-m)}$ such that

$$P_1 \begin{bmatrix} B_{21} \\ D \end{bmatrix} = \begin{bmatrix} B_2 \\ 0 \end{bmatrix}$$

where $B_2 \in \mathbb{R}^{(q-m) \times s}$ has full row rank with

$$\gamma = \text{rank} \begin{bmatrix} E & B \\ 0 & D \end{bmatrix}.$$

Write

$$P_1 \begin{bmatrix} A_{21} & A_{22} & B_{21} \\ C_{11} & C_{22} & D \end{bmatrix} = \begin{bmatrix} A_2 & A_3 & B_2 \\ A_4 & A_5 & 0 \end{bmatrix} \quad (9)$$

$\tilde{P} = \text{diag}(I_r, P_1) \text{diag}(\tilde{P}, I_m)$ and $\tilde{Q} = \text{diag}(\tilde{Q}, \tilde{Q}, I_s)$. In view of (8) and (9), we get

$$\tilde{P} \begin{bmatrix} E & A & B \\ 0 & C & D \end{bmatrix} \tilde{Q} = \begin{bmatrix} I_r & 0 & A_{11} & A_{12} & B_{11} \\ 0 & 0 & A_2 & A_3 & B_2 \\ 0 & 0 & A_4 & A_5 & 0 \end{bmatrix}. \quad (10)$$

This means that the system (1), (2) can be rewritten as

$$\dot{\xi}_1 = A_{11}\xi_1 + A_{12}\xi_2 + B_{11}u - P_{1,2}y \quad (11)$$

$$0 = A_2\xi_1 + A_3\xi_2 + B_2u - P_{2,2}y \quad (12)$$

$$0 = A_4\xi_1 + A_5\xi_2 - P_{3,2}y \quad (13)$$

where $\xi = [\xi_1^T \ \xi_2^T \ \xi_3^T]^T = \tilde{Q}^T x$; $[P_{1,2}^T \ P_{2,2}^T \ P_{3,2}^T]^T$ is the last $(q+m)$ columns of \tilde{P} . Due to the special structure of \tilde{P} , it is easy to verify that $P_{12} = 0$. The row partition of the last $(q+m)$ columns of \tilde{P} is obvious according to (11)–(13).

Since

$$\text{diag}(\tilde{P}, P_1) \begin{bmatrix} E & A & B \\ 0 & C & D \\ 0 & E & 0 \end{bmatrix} \tilde{Q} = \begin{bmatrix} I_r & 0 & A_{11} & A_{12} & B_{11} \\ 0 & 0 & A_2 & A_3 & B_2 \\ 0 & 0 & A_4 & A_5 & 0 \\ 0 & 0 & I_r & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (14)$$

and

$$\tilde{P} \begin{bmatrix} E & B \\ 0 & D \end{bmatrix} \text{diag}(\tilde{Q}, I_s) = \begin{bmatrix} I_r & 0 & B_{11} \\ 0 & 0 & B_2 \\ 0 & 0 & 0 \end{bmatrix} \quad (15)$$

it is easy to verify that the condition (7) holds if and only if A_5 has full column rank $n-r$.

Solving (13) for ξ_2 , substituting the solution into (11) and (12), and premultiplying (13) by $(I_\beta - A_5 A_5^+)$, where $\beta = q+m-r$, we get

$$\dot{\xi}_1 = \bar{A}_1 \xi_1 + B_{11}u + G_1 y \quad (16)$$

$$\bar{y} = \bar{C}_1 \xi_1 + \bar{B}_2 u \quad (\bar{y} = Hy) \quad (17)$$

$$\xi_2 = -A_5^+ A_4 \xi_1 + A_5^+ P_{3,2} y \quad (18)$$

where $\bar{A}_1 = A_{11} - A_{12} A_5^+ A_4$, $G_1 = A_{12} A_5^+ P_{3,2}$, $A_5^+ = (A_5^T A_5)^{-1} A_5^T$,

$$\bar{C}_1 = \begin{bmatrix} A_2 - A_3 A_5^+ A_4 \\ (I_\beta - A_5 A_5^+) A_4 \end{bmatrix}, \quad \bar{B}_2 = \begin{bmatrix} B_2 \\ 0 \end{bmatrix}.$$

$$H = \begin{bmatrix} P_{2,2} - A_3 A_5^+ P_{3,2} \\ (I_\beta - A_5 A_5^+) P_{3,2} \end{bmatrix}. \quad (19)$$

Now it is clear if the pair $\{\bar{A}_1, \bar{C}_1\}$ is detectable, one can design a full-order observer for the subsystem described by (16), (17) using the standard techniques. Then, combining the observer with (18) and considering the relation $x = \tilde{Q}\xi$, one finishes the observer design. The observer has the form (3), (4) with the order r .

In the following, we prove that under the condition (7), the pair $\{\bar{A}_1, \bar{C}_1\}$ is detectable if and only if the condition (6) holds.

Consider only

$$\begin{aligned} \text{rank} \begin{bmatrix} -\lambda E + A \\ C \end{bmatrix} &= \text{rank} \begin{bmatrix} -\lambda I_r + A_{11} & A_{12} \\ A_2 & A_3 \\ A_4 & A_5 \end{bmatrix} \\ &= \text{rank} \begin{bmatrix} -\lambda I_r + A_{11} & A_{12} \\ A_2 & A_3 \\ (I_\beta - A_5 A_5^+) A_4 & 0 \\ A_5^+ A_4 & I_{n-r} \end{bmatrix} \\ &= \text{rank} \begin{bmatrix} -\lambda I_r + (A_{11} - A_{12} A_5^+ A_4) & 0 \\ A_2 - A_3 A_5^+ A_4 & 0 \\ (I_\beta - A_5 A_5^+) A_4 & 0 \\ A_5^+ A_4 & I_{n-r} \end{bmatrix} \end{aligned} \quad (20)$$

where the second equality is obtained by premultiplying the line block $[A_4 \ A_5]$ in the first equality by the column nonsingular matrix

$$\begin{bmatrix} I_\beta - A_5 A_5^+ \\ A_5^+ \end{bmatrix}. \quad \square$$

The above observer is of order r . Noting the special structure of \bar{B}_2 , one knows that the second part of the measurement \bar{y} in (17) is free of inputs. Hence, it is easy to design a reduced-order observer in the form (3), (4) for the subsystem (16), (17) following, for instance a similar technique in [23]. The resulting observer will have the order p

$$\begin{aligned} p &= r - \text{rank}(I_\beta - A_5 A_5^+) A_4 \\ &= r - \{\text{rank}[A_4 \ A_5] - \text{rank} A_5\} \\ &= r - \left\{ \text{rank} \begin{bmatrix} E & A & B \\ 0 & C & D \end{bmatrix} - \text{rank} \begin{bmatrix} E & B \\ 0 & D \end{bmatrix} - (n-r) \right\} \\ &= n + \text{rank} \begin{bmatrix} E & B \\ 0 & D \end{bmatrix} - \text{rank} \begin{bmatrix} E & A & B \\ 0 & C & D \end{bmatrix} \end{aligned} \quad (21)$$

which is directly obtainable by using the form (16)–(18).

In the next section we shall prove that the conditions (6), (7) and the order (21) are equivalent to those given in [8] provided $\lambda E - A$ is regular and $D = 0$. Nevertheless, the condition (7) combining with the condition (6) is still only sufficient for the design of observers

of the type (3), (4), and the order (21) is not minimal generally. These can be shown by constructing counterexamples. For instance, Example 2 in [10] serves as a counterexample showing nonnecessity of (7) for the design of observers of the type (3), (4).

III. RELATIONSHIP AMONG EXISTING DESIGN CONDITIONS

In [7] and [8], the condition corresponding to (7) is

$$\text{rank} \left\{ \begin{bmatrix} E \\ C \end{bmatrix} J^T \begin{bmatrix} I \\ \Lambda \end{bmatrix} \right\} = n \quad (22)$$

where $n = \text{rank}[F \ B]$, J is the permutation matrix such that

$$J(\mu F + I)^{-1}[F \ B] = \begin{bmatrix} F_I & B_I \\ E_{II} & B_{II} \end{bmatrix} \quad (23)$$

where $[E_I \ B_I]$ has full row rank n , and Λ is the unique solution of the equation $[E_{II} \ B_{II}] = \Lambda[E_I \ B_I]$, μ is an arbitrary number such that $\mu I + I$ is nonsingular. Obviously the matrix pencil $\Lambda F - I$ is assumed to be regular. Since the condition (7) does not imply this assumption and the matrix D in (7) is generally nonzero, the following proposition indicates that the condition (7) is a generalization of the condition (22).

Proposition 3.1 Provided the matrix pencil $\Lambda F - I$ is regular, the following conditions are equivalent

$$a) \text{rank} \left\{ \begin{bmatrix} I \\ C \end{bmatrix} J^T \begin{bmatrix} I \\ \Lambda \end{bmatrix} \right\} = \text{rank}[I \ B] \quad (24)$$

$$b) \text{rank} \left\{ \begin{bmatrix} I \\ C \end{bmatrix} (\mu F + I)^{-1} [I \ B] [F \ B] \right\} = \text{rank}[F \ B] \quad (25)$$

$$c) \text{rank} \begin{bmatrix} I & A & B \\ 0 & C & 0 \\ 0 & F & 0 \end{bmatrix} = n + \text{rank}[F \ B] \quad (26)$$

where the notation $(\)$ represents an arbitrary generalized inverse of $(\)$ satisfying $(\)(\)(\) = (\)$.

Proof First we prove the equivalence between a) and b).

Since

$$\begin{aligned} &\text{rank} \left\{ \begin{bmatrix} I \\ C \end{bmatrix} (\mu F + I)^{-1} [F \ B] [I^* \ B] \right\} \\ &= \text{rank} \left\{ \begin{bmatrix} I \\ C \end{bmatrix} J^T S S^{-1} J (\mu L + I)^{-1} \right. \\ &\quad \left. [F \ B] [I^* \ B] (\mu F + I)^T S \right\} \end{aligned} \quad (27)$$

for any permutation matrix J and any nonsingular matrix S , by choosing J as in (23) and

$$S = \begin{bmatrix} I & 0 \\ \Lambda & I \end{bmatrix}$$

one has

$$\begin{aligned} S^{-1} J (\mu F + I)^{-1} [F \ B] &= \begin{bmatrix} I_I & 0 \\ -\Lambda & I_{II} \end{bmatrix} \begin{bmatrix} F_I & B_I \\ E_{II} & B_{II} \end{bmatrix} \\ &= \begin{bmatrix} E_I & B_I \\ 0 & 0 \end{bmatrix} \end{aligned} \quad (28)$$

which leads to

$$\begin{aligned} S^{-1} J (\mu F + I)^{-1} [F \ B] [I^* \ B] (\mu F + I)^T S \\ &= \{ S^{-1} J (\mu F + I)^{-1} [F \ B] \} \{ S^{-1} J (\mu F + I)^{-1} [F \ B] \}^* \\ &= \begin{bmatrix} E_I & B_I \\ 0 & 0 \end{bmatrix} \begin{bmatrix} F_I & B_I \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I_I & 0 \\ 0 & 0 \end{bmatrix} \end{aligned} \quad (29)$$

This, in view of (27), confirms the equivalence between a) and b).

Now we prove the equivalence between b) and c). To verify

$$\begin{aligned} &\begin{bmatrix} S^{-1} J (\mu F + I)^{-1} & 0 & 0 \\ 0 & 0 & I_r \\ 0 & I_n & 0 \end{bmatrix} \begin{bmatrix} I & A & B \\ 0 & C & 0 \\ 0 & F & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} I_I & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & H_1 & H & 0 \end{bmatrix} \end{aligned}$$

where

$$H_1 = \begin{bmatrix} I \\ C \end{bmatrix} J^T \begin{bmatrix} I_I \\ \Lambda \end{bmatrix}, \quad H = \begin{bmatrix} F \\ C \end{bmatrix} J^T \begin{bmatrix} 0 \\ I \end{bmatrix}$$

Then, due to the full row rank of $[F_I \ B_I]$ the conclusion is obvious. \square

In order to design Luenberger observers, different conditions are adopted in various design techniques. As shown by Dai [5] the condition (6) is necessary (but generally not sufficient) for any Luenberger type observer which provides the estimation of the descriptor vector x . We need only to compare the conditions which are additionally required for designing Luenberger observers. In the following, we outline these existing additional conditions which are given in terms of the original system matrices, and for reference we cite some corresponding works in which these conditions are used to design Luenberger observers.

$$a) \text{rank} \begin{bmatrix} I \\ C \end{bmatrix} = n \quad (\text{see [2], [19], [21]})$$

$$b) \text{rank} \begin{bmatrix} F & A & B \\ 0 & C & 0 \\ 0 & E & 0 \end{bmatrix} = n + \text{rank}[I \ B]$$

(see this note and [7], [8])

$$c) \text{rank} \begin{bmatrix} F & A \\ 0 & C \\ 0 & E \end{bmatrix} = n + \text{rank } F \quad (\text{see [4], [9], [22]})$$

$$d) \text{rank} \begin{bmatrix} I & A & B & 0 \\ 0 & F & 0 & A \\ 0 & 0 & 0 & C \\ 0 & 0 & 0 & F \end{bmatrix} = n + \text{rank} \begin{bmatrix} I & A & B \\ 0 & I & 0 \end{bmatrix}$$

(see [10])

Remark 3.1 For the convenience of comparison, $D = 0$ is assumed in (2). Thus, the condition (7) now has the form b).

Remark 3.2 In a recent paper [11] the necessary and sufficient conditions are provided for the Luenberger observer design in terms of the Weierstrass form of a transformed descriptor system. Therefore, the underlying descriptor systems are assumed to be regular. As a matter of fact, in terms of the Kronecker form [13], the same can be done for arbitrary descriptor systems.

It is known that, provided the necessary condition (6) holds under either conditions a) or b), the class of Luenberger observers of the form (3)–(4) can be designed. But under neither c) nor d) can this be guaranteed. It is also known that a general Luenberger observer, i.e., in (4), the input term may exist, can be designed under each of the conditions a)–d) combining with the condition (6). Since all these conditions a)–d) with (6) are known to be sufficient for the corresponding cases, it is interesting to know what kind of relationship exists among these conditions. The following proposition gives the answer.

Proposition 3.2 There exists the relation $a) \Rightarrow b) \Rightarrow c) \Rightarrow d)$ but, in general, $a) \Leftarrow b) \Leftarrow c) \Leftarrow d)$. Here, for instance $a) \Rightarrow b)$ represents that b) can be deduced from a).

Proof: On account of (10), without loss of generality, the matrices E , A , B , and C are assumed to be in the forms

$$E = \begin{bmatrix} I_1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_2 & A_4 \\ A_4 & A_5 \end{bmatrix},$$

$$B = \begin{bmatrix} B_{11} \\ B_2 \\ 0 \end{bmatrix}, \quad C = [C_{11} \quad C_{12}] \quad (31)$$

where B_2 has full row rank. By using these forms it is easy to prove the conclusion. \square

Proposition 3.2 shows that d) is the weakest condition among a)–d).

Finally, we provide a conclusion about the order of the observers. This result is parallel to Proposition 3.1, and its proof is analogous to that of Proposition 3.1.

Proposition 3.3: Provided the matrix pencil $\lambda E - A$ is regular and $D = 0$, the reduced order (21) of the designed observer can be expressed in the following different ways:

$$a) \quad p = \text{rank} [E \quad B] - \text{rank} \left\{ C^T J^T \begin{bmatrix} I_2 \\ N \end{bmatrix} \right\} \quad (32)$$

$$b) \quad p = \text{rank} [E \quad B] - \text{rank} \{ C(\mu E + A)^{-1} [E \quad B] [E \quad B]^{-1} \} \quad (33)$$

$$c) \quad p = \text{rank} [E \quad B] - \text{rank} \begin{bmatrix} E & A & B \\ 0 & C & 0 \end{bmatrix} + n. \quad (34)$$

Remark 3.3: The expression (32) was given in [8].

IV. CONCLUSION

A class of Luenberger observers in [7] and [8] can easily be designed under two simple conditions. The conditions in terms of the original system matrices are shown to be equivalent to those in [7] and [8]. Furthermore, the present design is available for the more general class of linear descriptor systems. However, in contrast with the claim in [7] and [8], the design conditions are known to be sufficient in usual cases. The exact relationship among the different design conditions in literature which are given in terms of the original system matrices is shown.

ACKNOWLEDGMENT

The authors appreciate the inspired comments of a reviewer on uniqueness of trajectories of arbitrary descriptor systems.

REFERENCES

- [1] M. El-Tohami, V. Lovass-Nagy, and R. Mukundan, "On the design of observers for generalized state space systems using singular value decomposition," *Int. J. Contr.*, vol. 38, pp. 673–683, 1983.
- [2] M. H. Verhaegen and P. Van Dooren, "A reduced order observer for descriptor systems," *Syst. Contr. Lett.*, vol. 8, pp. 29–37, 1986.
- [3] M. El-Tohami, V. Lovass-Nagy, and R. Mukundan, "Design of observers for time-varying discrete-time descriptor systems," *Int. J. Contr.*, vol. 46, pp. 841–848, 1987.
- [4] L. Dai, "Observers for discrete singular systems," *IEEE Trans. Automat. Contr.*, vol. AC-33, pp. 187–191, 1988.
- [5] L. Dai, *Singular Control Systems*. Berlin: Springer, 1989.
- [6] L. F. Lewis, "Geometric design techniques for observers in singular systems," *Automatica*, vol. 26, pp. 411–415, 1990.
- [7] P. N. Paraskevopoulos and F. N. Koumboulis, "Unifying approach to observers for regular and singular systems," *IEE Proc. D*, vol. 138, pp. 561–572, 1991.
- [8] —, "Observers for singular systems," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1211–1215, 1992.
- [9] D. N. Shields, "Observers for descriptor systems," *Int. J. Contr.*, vol. 55, pp. 249–256, 1992.
- [10] P. C. Müller and M. Hou, "On the observer design for descriptor systems," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 1666–1671, 1993.
- [11] A. Ailon, "On the reduced-order causal observer design for generalized control systems," *Int. J. Contr.*, vol. 57, pp. 1311–1323, 1993.
- [12] J. D. Cobb, "Controllability, observability, and duality in singular systems," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 1076–1082, 1984.
- [13] F. R. Gantmacher, *The Theory of Matrices, Vol. II*. New York: Chelsea, 1959.
- [14] L. R. Fletcher, "Regularizability of descriptor systems," *Int. J. Syst. Sci.*, vol. 17, pp. 843–847, 1986.
- [15] P. N. Paraskevopoulos, F. N. Koumboulis, K. G. Tzarakis, and G. E. Panagiotakis, "Observer design for generalized state space systems with unknown inputs," *Syst. Contr. Lett.*, vol. 18, pp. 309–321, 1992.
- [16] M. Hou and P. C. Müller, "Comments on 'Unifying approach to observers for regular and singular systems,'" *IEE Proc. D*, vol. 140, pp. 140–142, 1993.
- [17] P. N. Paraskevopoulos and F. N. Koumboulis, "Author's reply," *IEE Proc. D*, vol. 140, pp. 142–144, 1993.
- [18] V. C. Klema and A. J. Laub, "The singular value decomposition: Its computation and some applications," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 164–176, 1980.
- [19] B. Shafai and R. L. Carroll, "Design of a minimal-order observer for singular systems," *Int. J. Contr.*, vol. 45, pp. 1075–1081, 1987.
- [20] N. Minamide, N. Arii, and Y. Uetake, "Design of observers for descriptor systems using a descriptor standard form," *Int. J. Contr.*, vol. 50, pp. 2141–2149, 1989.
- [21] Y. Uetake, "Pole assignment and observer design for continuous descriptor systems," *Int. J. Contr.*, vol. 50, pp. 89–96, 1989.
- [22] S. Kawaji, "Design of observer for linear descriptor systems," in *Proc. IFAC World Cong.*, vol. 2, Tallinn, 1990, pp. 202–206.
- [23] C. T. Chen, *Linear System Theory and Design*. New York: Holt Rinehart, Winston, 1984.

Comments on "On the Robust Popov Criterion for Interval Lur'e Systems"

T. Mori, T. Nishimura, Y. Kuroe, and H. Kokame

Abstract—This correspondence supplements the remarks made in a paper by Dahleh *et al.* as to the Popov sector of Lur'e systems with interval plants. A numerical example is provided to show that strict inequalities hold among three sectors: those obtained by the theorem of the paper,¹ by applying Popov theorem to each member system, and by applying that to each of the systems with Kharitonov plants. This counterexample negates a tempting conjecture on "strong" robust Popov criterion.

In the above paper,¹ Dahleh, Tesi and Vicino derived the robust Popov criterion for a Lur'e system including a sector-type nonlinearity and a family of interval plants. Hereafter, we mostly employ

Manuscript received March 16, 1994.

T. Mori, T. Nishimura, and Y. Kuroe are with the Department of Electronics and Information Science, Kyoto Institute of Technology, Matsugasaki, Sakyo, Kyoto 606, Japan.

H. Kokame is with the Department of Electrical Engineering, Osaka Institute of Technology, Ohmiya, Asahi, Osaka 535, Japan.

IEEE Log Number 9407027.

¹M. Dahleh, A. Tesi, and A. Vicino, *IEEE Trans. Automat. Contr.*, vol. 38, pp. 1400–1405, Sept. 1993.

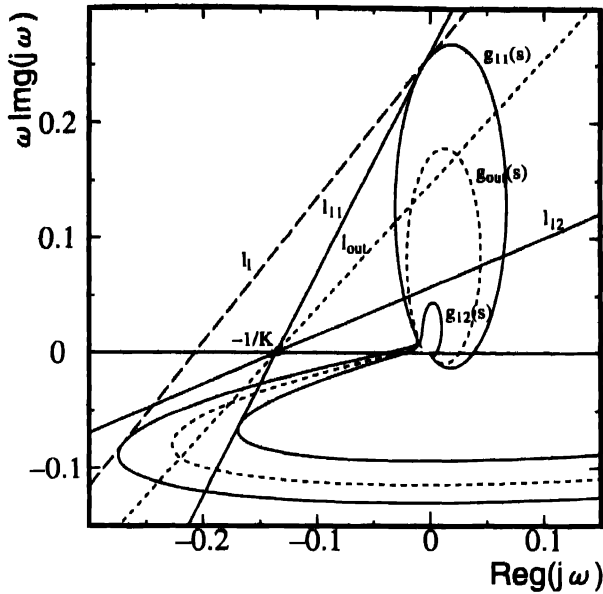


Fig. 1. Popov loci and Popov lines for the interval plant family

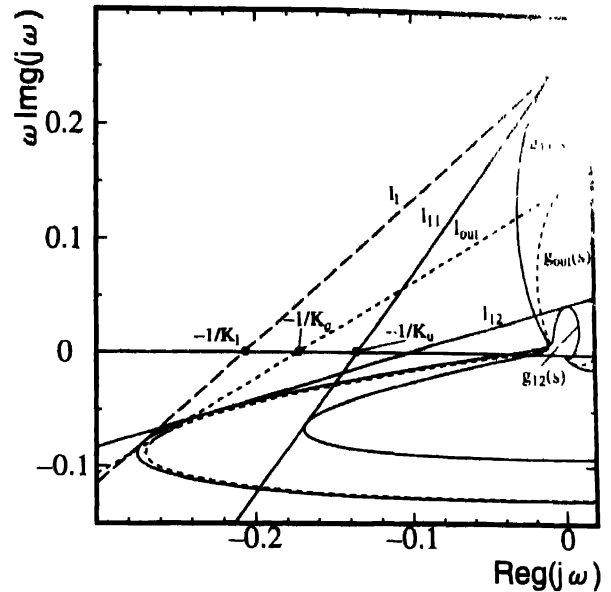


Fig. 2. Critical Popov lines and corresponding sectors.

the notations of the paper.¹ In particular, by "sector k ," we mean the sector $[0, k]$. In short, the criterion says that we have only to test absolute stability with Popov theorem for 16 Lur'e systems that correspond to Kharitonov plants. The caveat is, however, that we should choose a fixed multiplier constant θ_l for all these systems. This leads to a lower bound k_l for the sector k_0 which is obtainable by applying the Popov theorem to each member Lur'e system with the multiplier depending upon the member plant. If we need to apply the Popov theorem to only the 16 extreme systems with possibly different multipliers, then we have the sector k_u which is larger than or equal to k_0 . Thus, we have, in general, $k_l \leq k_0 \leq k_u$, as remarked in the paper.¹

Here, what we are intrigued by is the question: can the last inequality always be equality? If so, we would be able to obtain the "strong" robust Popov criterion, which we can afford most using the original Popov theorem. This tempting conjecture was also posed in [1]. Unfortunately, a counterexample has been found by an extensive search.

Setting $C(s) \equiv 1$, as in the paper,¹ we consider a stable interval plant,

$$g(s) = \frac{0.01s + 1}{s^5 + 5s^4 + a_1s^3 + 41s^2 + a_1s + 1} \quad (1)$$

where $a_1 \in [8, 11]$, $a_3 \in [10, 12]$. For this family, we have the following two Kharitonov plants:

$$\begin{aligned} g_{11}(s) &= g_{14}(s) = N_1(s)/D_1(s) = N_1(s)/D_1(s) \\ &= \frac{0.01s + 1}{s^5 + 5s^4 + 10s^3 + 41s^2 + 11s + 1} \end{aligned} \quad (2)$$

$$\begin{aligned} g_{12}(s) &= g_{13}(s) = N_1(s)/D_2(s) = N_1(s)/D_3(s) \\ &= \frac{0.01s + 1}{s^5 + 5s^4 + 12s^3 + 41s^2 + 8s + 1} \end{aligned} \quad (3)$$

Other than these plants, we pick a non-Kharitonov plant

$$g_{00}(s) = \frac{0.01s + 1}{s^5 + 5s^4 + 10s^3 + 41s^2 + 9s + 1} \quad (4)$$

Fig. 1 shows the Popov loci for $g_{11}(s)$ and $g_{12}(s)$ (using a solid curve) and that for $g_{00}(s)$ (using a dashed curve). We can draw the line l_1 touching both of the Kharitonov loci, and this determines the sector obtained by "weak" robust Popov criterion. For $k = K = 7.363 \dots$ (corresponds to $-1/K$ in Fig. 1), the criterion is no more effective, but we can still draw the lines, l_{11} and l_{12} , in whose right-hand side the Kharitonov loci are located. However, it is impossible to draw the Popov line for g_{00} at the point $(-1/K, 0)$. This indicates that the "strong" robust Popov criterion fails to hold.

In order to obtain k_0 , we have to resort to partitioning the intervals into evenly spaced grids. For each coefficient value at the grid, the existence of the multiplier is examined. For this task, a symbolic computation approach is available [2]. Fig. 2 illustrates three critical Popov lines for the above interval plant. These lines determine the Popov sectors: $k_l = K_l = 4.826 \dots$, $k_0 = K_0 = 5.774 \dots$, and $k_u = K_u = 7.363 \dots$. The plant $g_{00}(s)$ which gives K_0 in this figure is slightly different from $g_{00}(s)$ in Fig. 1, in that the coefficient 9 of the first degree term in the denominator in (4) is changed to 8. We see, from this example, that the inequalities among the sectors can be actually strict: $k_l < k_0 < k_u$.

ACKNOWLEDGMENT

The authors are indebted to J. Ogiwara for his exhaustive computation.

REFERENCES

- [1] T. Mori, "Strong Popov theorem for Lur'e systems with interval plants," in *Robustness of Dynamical Systems with Parameter Uncertainties*, M. Mansour, et al., Eds. Basel: Birkhaeuser, 1992, p. 313.
- [2] R. Wang, "Symbolic computation approach for absolute stability," *Int. J. Contr.*, vol. 58, no. 2, pp. 495-502, 1993.

An Output Feedback Globally Stable Controller for Induction Motors

Gerardo Espinosa-Pérez and Romeo Ortega

Abstract—We present here a globally stable nonlinear dynamic output feedback controller for torque tracking and flux regulation of induction motors. The control law is globally defined (even in startup), requires only measurement of stator variables and rotor speed, and does not rely on cancellation of the systems nonlinearities. Our work extends in several directions the result of the paper by Ortega *et al.*, where the torque tracking problem was solved for a model where the variables are expressed in a frame rotating at an arbitrary angular frequency (dq model). First, we obviate the need to transfer the dq control signals of the paper by Ortega *et al.* to the physical input variables in the stator frame, hence providing a directly implementable control law. Second, besides the torque tracking objective, we include the practically important rotor flux regulation task. Third, by choosing a more suitable representation of the motor model, we simplify the controller structure and provide a better understanding of its derivation and behavior.

I. INTRODUCTION

It is widely recognized that induction motors, which have been extensively utilized in industrial controls, offer a very promising alternative for future applications like electrical vehicles and robotics, see, e.g., [21] and [1]. Among the advantages of the ac induction motor over dc types: they pack more power into less space, they have no need for failure-prone brushes, they are more rugged, and they have flatter efficiency-versus-speed curves. Also, with the continuous drop in the cost of electronics, the ac system is becoming more cost-effective. To match the higher quality standards of wider drive ranges and higher reliability requires the development of new robust control techniques—a challenging task that has attracted the attention of control theoreticians.

Induction motors can be described [22] by a fifth-order nonlinear differential equation with four electrical coordinates (stator and rotor currents or fluxes) and one mechanical coordinate (rotor speed). There are two physical inputs to the motor (stator voltages), and only three state variables are typically available for measurement (stator currents and the rotor speed). Furthermore, the motor parameters vary considerably during operation and the generated torque is perturbed by an unknown torque load. In typical industrial drives, the controlled variables are rotor position and speed, which are indirectly regulated via nested PI loops around the fundamental torque control loop [11]. With an eye on electric vehicles and robotic applications, we center here directly in torque tracking.¹

A globally stable output feedback controller for torque regulation of a complete induction motor model with unknown torque load was recently presented in [16]. These results were further extended to torque tracking in [17]. The control algorithms in both papers are derived based on the so called dq model of the motor. This model is obtained applying two coordinate transformations to the basic $\alpha\beta$

model. First, we refer to rotor variables to the stator coordinate frame² to obtain the so-called ab model. Then, an additional projection of all signals to an arbitrarily rotating frame is applied. In the industrial electronics community, where the focus is on steady-state operation at constant speed, these transformations are well understood, see, e.g., [8], [12]. However, as pointed out in Section II-B, there seems to be some confusion among the control workers as to how these dynamical models and corresponding controllers are related. One of our objectives in writing this paper is to clarify these relationships.

Since the controllers in [16] and [17] are derived for the dq model, an additional step to transfer the control signals to the physical input variables in the stator frame is needed for its implementation. In this note we show that, contrary to some incorrect statements made in the literature [13], this transformation does not require the measurement of rotor variables. To enhance the readability of the note, instead of deriving the implementable control law “reverse transforming” the control of [17], we directly obtain the control in the $\alpha\beta$ model. We then show how the scheme of [17] can be obtained by transforming the present control law. In summary, our main contribution is then to provide the first directly implementable³ globally stable nonlinear dynamic output feedback controller for torque tracking and flux regulation of an induction motor.

The difficulty of torque control is that, in contrast with dc motors, torque is a nonlinear function of fluxes and currents. It has been shown in [3] that the full induction motor model (including the mechanical dynamics) is not exactly linearizable if we select torque as the output. It is, however, input-output linearizable (modulo singularity problems) with respect to velocity and rotor flux norm [9]. This theoretically interesting property has motivated several authors to study the velocity tracking problem, e.g., [10], [9], [3], [14], [2]. In [18] it is shown how to achieve global speed regulation from our torque tracking controller. It is worth pointing out that, in contrast with our control law which is globally defined, even in start up, in the aforementioned results (see also [6]) various singularity problems appear.

The remainder of the note is organized as follows. First, we derive in Section II the induction motor basic $\alpha\beta$ model, establish some key properties, and present the torque tracking and flux regulation problem. Section III contains the proposed control law, derived directly from the $\alpha\beta$ model, hence implementable without further transformations. The proof of the main stability result is deferred to the Appendix. In Section IV we explain the different transformations of the motor model, and using these transformations rederive the controller of [17]. The performance of the proposed algorithm is illustrated via simulations in Section V. Finally, we present some concluding remarks.

II. MODEL AND PROBLEM FORMULATION

In this section we derive first the basic motor model ($\alpha\beta$ model) from the total energy function and Lagrange's equations, and establish its key *passivity* property.⁴ Also, we present some fundamental relationships of the motor that are used in the sequel for the controller derivation.

²This is a classical approach that simplifies the analysis of the machine removing from the electrical equations the dependence on position.

³To obtain the 3ϕ equivalent controller, it is only necessary to apply the inverse Blondel's transformation [15].

⁴It is clear that passivity, being an input-output property, is independent of the system coordinates thus valid in all motor representations.

Manuscript received May 7, 1993; revised August 25, 1993, November 16, 1993, and March 21, 1994. The work of R. Ortega was supported in part by the Commission of European Communities under Contract ERB CHRX CT 93-0380. Part of this work was carried out while the first author was with DEPEI-UNAM.

G. Espinosa-Pérez is with the Universidad Nacional de México, Instituto Ingeniería, México.

R. Ortega is with the Université de Compiègne, URA CNRS 817, PB 649, 60206 Compiègne cedex, France.

IEEE Log Number 9407028.

¹See Remark 2.1 for the case of speed/position control.

A Basic $\alpha\beta$ Model

We consider a standard three phase single pole pair induction motor represented with a two-phase model known as the $\alpha\beta$ model⁵.

It considers two coils for the stator and two coils for the rotor in pairs in two orthogonal $\alpha\beta$ axes. The axes for the stator have a fixed position, while those corresponding to the rotor are rotating at the rotor angular velocity q . We make the usual basic assumptions about the mechanical structure, balanced energy distribution and linear relationship between the flux vector $\lambda = [\lambda_s^T \lambda_r^T]^T = [\lambda_1 \lambda_2 \lambda_3 \lambda_4]^T \in \mathbb{R}^4$ and the current vector $q = [q_s^T q_r^T]^T = [q_1 q_2 q_3 q_4]^T \in \mathbb{R}^4$

$$\lambda = D(q)q \quad (2.1)$$

where the subscripts (\cdot) , (\cdot) , (\cdot) are used throughout the note to denote electrical, stator, and rotor variables respectively. The superscript $(\cdot)^T$ denotes transposition. $D(q) = D^T(q) > 0$ is the 4×4 inductance matrix of the windings given by

$$D(q) = \begin{bmatrix} I_s & I_m & I_s e^{j\theta} \\ I_s e^{-j\theta} & I_r & I_m \\ I_m & I_m & I_r \end{bmatrix}$$

where $I_s, I_r, I_m > 0$ are the stator, rotor, and mutual inductances respectively. I is the 2×2 identity matrix and $e^{j\theta}$ is the rotation matrix

$$e^{j\theta} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

with J the 2×2 antisymmetric matrix

$$J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

If we neglect shaft compliance, the generalized coordinates are the electrical charges in each of the coils q and the angular position of the rotor q . Since under this assumption there is no potential energy, the Lagrangian function coincides with the kinetic and the total energies and is given by

$$\mathcal{L}(q, \dot{q}) = \frac{1}{2} \dot{q}^T D(q) \dot{q} = \frac{1}{2} \dot{q}^T \begin{bmatrix} D(q) & 0 \\ 0 & D \end{bmatrix} \dot{q} \quad (2.2)$$

where $q = [q^T \dot{q}^T]^T \in \mathbb{R}^5$ and $D > 0$ is the rotor inertia.

Applying Euler-Lagrange equations, we get the state equations of the $\alpha\beta$ model describing both the electrical and mechanical subsystems as

$$D(q)\dot{q} + W_1(q)\dot{q}q + Rq = Mu \quad (2.3)$$

$$D\ddot{q} - \frac{1}{2}q^T W_1(q)\dot{q} + Rq = -\tau_l \quad (2.4)$$

where $R = \text{diag}\{R_s, R_s, R_r, R_r\} > 0$, $M = [I_2 \ 0]^T$, $W_1(q) = \partial D(q)/\partial q$, the control signals $u = [u_1 \ u_2]^T$ are the stator voltages, $R_s, R_r > 0$ are the stator and rotor resistances, $R > 0$ is the motor damping coefficient and τ_l is the load torque. The regulated outputs are the generated torque and the rotor flux which are given by

$$\tau = \frac{1}{2}q^T W_1(q)\dot{q}, \quad \lambda = L e^{-j\theta} q_s + I_r q_r \quad (2.5)$$

B Torque Tracking and Flux Regulation Problem

We consider the induction motor model (2.3)–(2.4) with outputs stator currents q_s and rotor speed \dot{q} . Control is subject to disturbance⁶ τ_l and regulated signals τ and λ . The torque reference τ_l is a bounded differentiable function with known first derivative. Our objective is then to design a control law that will ensure global asymptotic torque and flux regulation, that is $\lim_{t \rightarrow \infty} (\tau - \tau_l) = 0$, $\lim_{t \rightarrow \infty} \|\lambda - \lambda_d\| = 0$ for all initial conditions and with all internal signals uniformly bounded. Furthermore, we require that the motor in steady state operates in a balanced regime, i.e., sinusoidal stator voltage with three phases (one shifted 90° with respect to the other) with equal amplitude and frequency.

Remark 2.1 In [18] it is shown how from the torque tracking objective above we can solve the speed/position tracking problem. It essentially amounts to selecting the desired torque as a function of the desired speed.

C Motor Properties

1) Passivity. The proposition below reveals the fundamental energy dissipation property of the induction motor.

Proposition 2.1 (Passivity). The motor model (2.3)–(2.4) with zero load torque defines an output strictly passive operator $u \mapsto q$ that is there exists $\epsilon_1 > 0$, $\epsilon_2 \in \mathbb{R}$ such that $\int_0^t q^T u d\tau \geq \epsilon_1 \int_0^t \|q_s\|^2 d\tau + \epsilon_2$ (with $\|\cdot\|$ the Euclidean norm) holds for all $t > 0$ and all locally square integrable stator voltages. Furthermore, the workless forces may be factored as

$$\begin{bmatrix} W_1(q)\dot{q}q \\ -\frac{1}{2}q^T W_1(q)\dot{q} \end{bmatrix} = C(q, \dot{q})\dot{q} \\ = \begin{bmatrix} 0 & 0 & I J e^{j\theta} I \\ -L J e^{j\theta} I e^{-j\theta} q & 0 & 0 \\ I q^T J e^{j\theta} I e^{-j\theta} q & 0 & 0 \end{bmatrix} \dot{q}$$

where $D(q) = C(q, \dot{q}) + C^T(q, \dot{q})$.

Proof. Taking the time derivative of the system's total energy we get

$$\begin{aligned} \dot{\mathcal{L}} &= \dot{q}^T \left[\frac{1}{2} D(q)\dot{q} + \begin{bmatrix} -W_1(q)\dot{q}q \\ \frac{1}{2}q^T W_1(q)\dot{q} \end{bmatrix} - Rq + Mu + \xi \right] \\ &= \dot{q}^T (-Rq + Mu + \xi) \end{aligned}$$

where $R = \text{diag}\{R_s, R_s, R_r, R_r\}$, $M = [M^T \ 0]^T$, $\xi = [0 \ -\tau_l]^T$. From the integration of the last equation we get the energy balance equation

$$\underbrace{\mathcal{L}(t) - \mathcal{L}(0)}_{\text{stored energy}} = - \underbrace{\int_0^t \dot{q}^T Rq d\tau}_{\text{dissipated energy}} + \underbrace{\int_0^t \dot{q}^T (Mu + \xi) d\tau}_{\text{input energy}}$$

Noting that $\mathcal{L}(t) \geq 0$ and setting $\epsilon_1 = \lambda_{\min}\{R\}$ and $\epsilon_2 = -\mathcal{L}(0)$ proves the first part of the proposition. The second part of the proof is established via direct substitution of the terms. $\square\square\square$

The importance of the passivity property for our tracking problem is best appreciated with the following corollary that will be instrumental for the subsequent stability analysis. Its proof follows trivially from the derivative of the desired total energy function $\mathcal{L}_l = \frac{1}{2}C^T D(q)\dot{q}$ and the skew symmetry property

⁶For the sake of brevity we restrict our attention to the case of known disturbance. The extension to the unknown but linearly parameterized disturbance case follows verbatim from [17].

⁵For further details and generalizations see [15] and [4].

Corollary 2.1: For all locally square integrable q_s, \dot{q}_s and all $\mathcal{K} \in \mathcal{R}^{5 \times 5}$, the system

$$\mathcal{D}(q_s)\dot{e} + [\mathcal{C}(q_s, \dot{q}_s) + (\mathcal{R} + \mathcal{K})]e = \psi \quad (2.6)$$

satisfies the dissipation inequality $\int_0^t e^T \psi d\tau \geq c_3 \int_0^t \|e\|^2 d\tau + c_4 \|e(t)\|^2 - c_5 \|e(0)\|^2$, where $c_3 = \lambda_{\min}\{\mathcal{R} + \mathcal{K}\}$, $c_4 = \lambda_{\min}\{\mathcal{D}(q_s)\} > 0$, $c_5 = \lambda_{\max}\{\mathcal{D}(q_s)\} > 0$. Consequently, if $\psi \equiv 0$ and $\mathcal{R} + \mathcal{K} > 0$, then $\lim_{t \rightarrow \infty} e = 0$. $\square\square\square$

Remark 2.2: The factorization of the workless forces is not unique, but there always exists a suitable selection of $\mathcal{C}(q, \dot{q})$ to enforce the key skew-symmetry property [20]. It should be remarked that in the factorization of Proposition 2.1, the third and fourth row of $\mathcal{C}(q, \dot{q})$ are independent of \dot{q}_r . This feature will be used in the sequel.

2) Torque and Rotor Flux: In the remaining part of this subsection, we derive a key relationship between the rotor flux λ_r and the generated torque τ . To this end, we first notice from (2.3) that the rotor flux vector given in (2.1) satisfies the differential equation

$$\dot{\lambda}_r + R_r \dot{q}_r = 0. \quad (2.7)$$

Now, from the definition of rotor flux angle $\rho = \arctan\{\lambda_1/\lambda_3\}$, we have

$$\dot{\rho} = \frac{-1}{\|\lambda_r\|^2} \lambda_r^T \mathcal{J} \dot{\lambda}_r = \frac{R_r}{\|\lambda_r\|^2} \lambda_r^T \mathcal{J} \dot{q}_r,$$

where the last equation was obtained using (2.7). Noting that $\tau = \lambda_r^T \mathcal{J} \dot{q}_r$, we establish the fundamental fact that the rotor flux vector rotates at an angular speed

$$\dot{\rho} = \frac{R_r}{\|\lambda_r\|^2} \tau. \quad (2.8)$$

Our interest in presenting the relations above stems from the fact that we can express the torque tracking and flux regulation problem in terms of regulation of rotor current and flux as shown below.

Proposition 2.2 (Rotor Flux Regulation): Given a smooth bounded function $\tau_d(t)$ and a positive constant $\beta > 0$, let

$$\lambda_{r,d} := \beta \begin{bmatrix} \cos(\rho_d) \\ \sin(\rho_d) \end{bmatrix}, \quad \dot{\rho}_d := \frac{R_r}{\beta^2} \tau_d, \quad \rho_d(0) = 0. \quad (2.9)$$

Assume $\lim_{t \rightarrow \infty} \lambda_r = \lambda_{r,d}$ and $\lim_{t \rightarrow \infty} \dot{\lambda}_r = \dot{\lambda}_{r,d}$, then $\lim_{t \rightarrow \infty} \|\lambda_r\| = \beta$, $\lim_{t \rightarrow \infty} \tau = \tau_d$. $\square\square\square$

III. CONTROLLER DESIGN

From Proposition 2.2 and (2.7), we see that the control objective is attained if we ensure that the rotor flux converges to the solution of

$$\dot{\lambda}_{r,d} = \frac{R_r \tau_d}{\beta^2} \mathcal{J} \lambda_{r,d}, \quad \lambda_{r,d}(0) = \begin{bmatrix} \beta \\ 0 \end{bmatrix} \quad (3.1)$$

and \dot{q}_r converges to

$$\dot{q}_{r,d} := -\frac{\tau_d}{\beta^2} \mathcal{J} \lambda_{r,d}. \quad (3.2)$$

Using (2.1), we can express these two conditions in terms of currents by requiring \dot{q}_s to converge to

$$\dot{q}_{s,d} := \left(\frac{1}{L_s} \mathcal{I}_2 + \frac{L_r}{L_s \beta^2} \tau_d \mathcal{J} \right) e^{\mathcal{J} q_s} \lambda_{r,d}. \quad (3.3)$$

In other words, $\dot{q}_{s,d}, \dot{q}_{r,d}$ above define a *desired behavior* for the motor currents that delivers the required torque and insures rotor flux regulation.

On the other hand, if we define the difference between the actual and the desired behavior as $e := [e_1^T, e_2^T, e_3^T]^T := \dot{q} - \dot{q}_d$, from Corollary 2.1 we see that if (for the given $\dot{q}_{s,d}, \dot{q}_{r,d}$) we can find $\dot{q}_{s,d}$ and u such that the following system is satisfied for all q_s, \dot{q}_s :

$$\mathcal{D}(q_s)\ddot{q}_d + \mathcal{C}(q_s, \dot{q}_s)\dot{q}_d + \mathcal{R}\dot{q}_d = \mathcal{M}u + \xi \quad (3.4)$$

with $\dot{q}_d := [\dot{q}_{s,d}^T, \dot{q}_{r,d}^T, \dot{q}_{s,d}^T]^T$, then we have convergence to the desired current behavior.

It is easy to see that this is achieved with (3.2), (3.3), and

$$u = L_s \ddot{q}_{s,d} + L_s e^{\mathcal{J} q_s} \ddot{q}_{r,d} + L_s \mathcal{J} e^{\mathcal{J} q_s} \dot{q}_{r,d} + R_s \dot{q}_{s,d} \quad (3.5)$$

$$D_m \ddot{q}_{s,d} + L_s \dot{q}_r^T \mathcal{J} e^{-\mathcal{J} q_s} \dot{q}_{s,d} + R_m \dot{q}_{s,d} = -\tau_L. \quad (3.6)$$

From Corollary 2.1, it can easily be shown that this control law ensures global asymptotic torque tracking and flux regulation; however, it requires the measurement of the rotor current. To overcome this problem, we propose below a scheme where we use instead its estimate, add to the desired "energy function" \mathcal{L}_d a term that depends on the observation errors, and use the damping term \mathcal{K} to achieve a sign definite derivative. These derivations are summarized in the proposition below.

Proposition 3.1 (Main Result). Consider the induction motor model (2.3), (2.4) with measurable outputs stator currents \dot{q}_s and rotor speed \dot{q}_s . Assume the load torque τ_L is a known bounded function and that the motor parameters are also known. The torque reference τ_d is a bounded differentiable function with known first derivative. Let the control law be defined as

$$u = \Gamma(\tau_d, \dot{\tau}_d) \mathcal{J} e^{-\mathcal{J} q_s} \lambda_{r,d} + L_s \mathcal{J} e^{-\mathcal{J} q_r} \dot{q}_{r,d} - K_1(\dot{q}_{s,d}) e_s \quad (3.7)$$

with

$$\begin{aligned} \Gamma(\tau_d, \dot{\tau}_d) &= \frac{L_s \sigma}{L_{sr} \beta^2} \dot{\tau}_d \mathcal{I}_2 + \frac{L_s}{L_{sr}} \left(\frac{L_r \tau_d}{\beta^2} \mathcal{J} + \mathcal{I}_2 \right) \\ &+ \frac{\tau_d}{L_{sr} \beta^2} \left(L_s \mathcal{I}_2 + \frac{L_r \sigma}{\beta^2} \mathcal{J} \right) R_r \\ &+ \frac{1}{L_{sr}} \left(\frac{\tau_d L_r}{\beta^2} \mathcal{I}_2 - \mathcal{J} \right) R_s \end{aligned} \quad (3.8)$$

$\sigma := L_s - L_{sr}^2/L_r$ and controller dynamics (3.1) and

$$\ddot{q}_{s,d} = -\frac{1}{D_m} [L_{sr} \dot{q}_r^T \mathcal{J} e^{-\mathcal{J} q_s} \dot{q}_{r,d} + R_m \dot{q}_{s,d} + \tau_L + K_2(\dot{q}_d) e_s];$$

$$\dot{q}_{s,d}(0) = \dot{q}_{s,d0}. \quad (3.9)$$

The gains $K_1(\dot{q}_{s,d})$ and $K_2(\dot{q}_d)$ are given as

$$K_1(\dot{q}_{s,d}) = \frac{L_{sr}^2}{4\epsilon} \dot{q}_{s,d}^2, \quad K_2(\dot{q}_d) = \frac{L_{sr}^2}{4\epsilon} (\dot{q}_{1,d}^2 + \dot{q}_{2,d}^2), \quad 0 < \epsilon < R, \quad (3.10)$$

while the state estimator is

$$D_s(q_s)\ddot{\hat{q}}_s + W_1 \dot{q}_s \dot{\hat{q}}_s + R_s \dot{\hat{q}}_s = \mathcal{M}u - L(q_s, \dot{q}_s) \dot{\hat{q}}_s \quad (3.11)$$

with $\dot{\hat{q}}_s := \dot{\hat{q}}_s - \dot{q}_s$ the observation error and

$$L(q_s, \dot{q}_s) = \begin{bmatrix} 0 & 0 \\ \mathcal{J}_2 e^{-\mathcal{J} q_s} & 0 \end{bmatrix} \dot{q}_s.$$

Under these conditions, the closed-loop system achieves global asymptotic torque tracking and flux norm regulation with balanced operation.

Remark 3.1 We should recall that we have posed here a *torque* control problem where the given desired torque automatically determines [via the relationship (2.4)] the rotor speed. In this sense q_{1i} is the rotor speed that corresponds to the desired torque. Notice from (3.6) that actually this is the case if q converges to q_i . In the case where we are given a rotor speed or position control objective, τ_i should be chosen as proposed in [18].

IV. COORDINATE TRANSFORMATIONS: *ab* AND *dq* MODELS

In this section, we introduce the coordinate transformations that give us the *ab* and *dq* models. Our objective in studying in detail these transformations is twofold: first, to clearly establish the relationship between the various models studied in the control literature, second, to prove that the control law of [17] can be obtained by transforming the present controller (or vice versa) and that this transformation does not require measurement of the rotor variables.

To obtain the *ab* model we follow the standard procedure of applying to the $\alpha\beta$ model a coordinate transformation that projects the rotor currents to the stator frame. This is achieved with the change of coordinates

$$q_{1i} = T_{1i}(q)q = \begin{bmatrix} I & 0 & 0 \\ 0 & e^{j\theta} & 0 \\ 0 & 0 & 1 \end{bmatrix} q$$

Further, to obtain the *dq* model we project the *ab* currents into a frame rotating at an arbitrary angular speed ω as [8]

$$q_{1i} = T_{1i}(q)q_{1i} = \begin{bmatrix} e^{j\theta} & 0 & 0 \\ 0 & e^{j\theta} & 0 \\ 0 & 0 & 1 \end{bmatrix} q_{1i}$$

where $\theta = \omega t + \theta(0) = 0$. The relationships between the various representations are easily defined with the general rotation matrix

$$T(I - \theta) = \begin{bmatrix} e^{j\theta} & 0 & 0 \\ 0 & e^{j\theta} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

which clearly satisfies $T(q - \theta) = T(0 - \theta)T(q - 0)$. Hence, $q_{1i} = T(0 - \theta)q_{1i} = T(q - \theta)q$. To obtain the inverse transformations notice that $T^{-1}(q - \theta) = T^{-1}(q - \theta)$.

Applying the latter transformation to the $\alpha\beta$ model and premultiplying the resulting equation by $T(q - \theta)$ yields, after some simple calculations, the *dq* model studied in [16] and [17]

$$D_{1i}q_{1i} + C_{1i}(q_{1i})q_{1i} + Rq_{1i} = Mu_{1i} + \xi \quad (4.1)$$

where $Mu_{1i} = T(q - \theta)Mu$, the matrices D_{1i} and C_{1i} are given by

$$D_{1i} = \text{diag}\{D^{1i}, D\} = \text{diag}\left\{\begin{bmatrix} I & I & I \\ I_s & I_s & L \\ L & L & I_s \end{bmatrix} D\right\} \in \mathbb{R}$$

$$C_{1i}(q_{1i}, \omega) = \begin{bmatrix} \begin{bmatrix} I & \mathcal{J} & I \\ I_s & \mathcal{J} & I \end{bmatrix} (\omega - q) & C_{1i}(q_{1i}) \\ -C_{1i}^T(q_{1i}) & 0 \end{bmatrix} \in \mathbb{R}^{n \times n}$$

$$C_{1i}(q_{1i}) = [(I - \mathcal{J}q^{1i} + I_s \mathcal{J}q^{1i})^T 0]^T$$

$$\text{and } q_{1i} = [(q^{1i})^T (q^{1i})^T q]^T$$

In order to show that the control of [17] can be obtained by means of the transformation presented above, notice that

$$u_{1i} = e^{-j\theta} u = I_s e^{-j\theta} q_{s1} + I_s e^{j(\theta_s - \theta)} q_{s2} + I_s \mathcal{J} e^{-j(\theta_s - \theta)} q_{s2} + R e^{-j\theta} q_{s1} - K_1(q_{s1})e^{-j\theta} \quad (4.2)$$

where $e^{-j(\theta_s - \theta)} q_{s1} = q_{s1}^{1i}$ and

$$e^{-j\theta} q_{s2} = \begin{bmatrix} -\frac{1}{\sqrt{2}} \frac{\tau_i}{R} (R - \tau_i + \frac{1}{2} q_{s1}) \\ \frac{1}{\sqrt{2}} (R - \tau_i + \frac{1}{2} q_{s1}) + \frac{1}{2} \end{bmatrix}$$

$$e^{-j(\theta_s - \theta)} q_{s2} = \begin{bmatrix} \frac{1}{\sqrt{2}} \frac{\tau_i}{R} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

$$e^{-j\theta} q_{s1} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \frac{\tau_i}{R} \end{bmatrix} \quad e^{-j(\theta_s - \theta)} q_{s1} = \begin{bmatrix} 0 \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

$$e^{-j\theta} e^{-j(\theta_s - \theta)} = \begin{bmatrix} q_{s1}^{1i} - \frac{1}{\sqrt{2}} \frac{\tau_i}{R} \\ q_{s2}^{1i} - \frac{1}{\sqrt{2}} \frac{\tau_i}{R} \end{bmatrix} \quad (4.3)$$

The last expressions were obtained considering that in this new reference frame the desired rotor flux will rotate at an angular speed $\theta = q_{s1}^{1i} e^{-j\theta} - \theta = q_{s1}^{1i}$ therefore $e^{-j(\theta_s - \theta)} e^{-j\theta} = [1 \ 0]^T$.

Substitution of the above expressions in (4.2) results in the following controller

$$u_{1i}^{1i} = -I_s q_{s1}^{1i} q_{s1}^{1i} + \frac{1}{2} - K_1^{1i} \left(q_{s1}^{1i} - \frac{1}{\sqrt{2}} \frac{\tau_i}{R} \right)$$

$$u_{1i}^{1i} = -I_s q_{s2}^{1i} q_{s2}^{1i} + \frac{1}{2} - K_1^{1i} \left(q_{s2}^{1i} - \frac{1}{\sqrt{2}} \frac{\tau_i}{R} \right)$$

where $K_1^{1i} = K_1$ and

$$\frac{1}{2} = \frac{R}{L} \frac{1}{2} - \frac{R}{L} (I - I_s - L) \tau_i - \frac{1}{L} \frac{1}{2} \tau_i q$$

$$= \frac{1}{L} \frac{1}{2} [(I - R + R I) \tau_i + (I - L - I_s) \tau_i + I \frac{1}{2} q]$$

The controller state is given by

$$q_{1i} = -\frac{1}{D} [I_s (q^{1i})^T \mathcal{J} q^{1i} + R q_{s1} + K_1^{1i} e^{-j\theta} + \tau_i]$$

with $K_1^{1i} = I_s / L e^{j\theta} [(q_{s1}^{1i})^2 + (q_{s2}^{1i})^2]$. The structure of the estimator state is easily obtained by a procedure similar to the one followed to obtain the model (4.1).

We must note the fact that the above controller differs slightly from the controller in [17] since the desired behavior for the currents (4.3) in this case was chosen to obtain the norm of the rotor flux equal to $\frac{1}{2}$ instead of $I_s \frac{1}{2}$ as in that paper.

Discussion. 1) From the *dq* model (4.1) we can obtain as a particular case the classical *ab* model by choosing $\omega = 0$.

2) A key observation is the introduction of ω in the model (4.1). This variable is used in [16] and [17] as an additional control input (denoted u_3 in those papers) to solve the torque control problem. From the derivations above, it is clear that given u_{1i} and ω we can obtain the *ab* controls via $Mu_{1i} = T^{-1}(q - \theta)Mu_{1i}$. The problem is that u_{1i} and ω themselves depend on variables expressed in the rotating frame. One objective of this note is to obviate the need of all these transformations by defining the controller dynamics directly in terms of the signals q_{s1}, q_{s2}, q .

3) It is important to remark that the *dq* model above is different from the field coordinate model used in field oriented control. The latter is obtained by selecting the angular frequency ω equal to the rotation angle of the flux vector, see [11, p. 216]. See also Section III A of [14] and [3]. It is clear that to obtain the physical signals from the field coordinate model, we require knowledge of the full state vector. This is, however, not the case for our *dq* model as erroneously pointed out in [13].

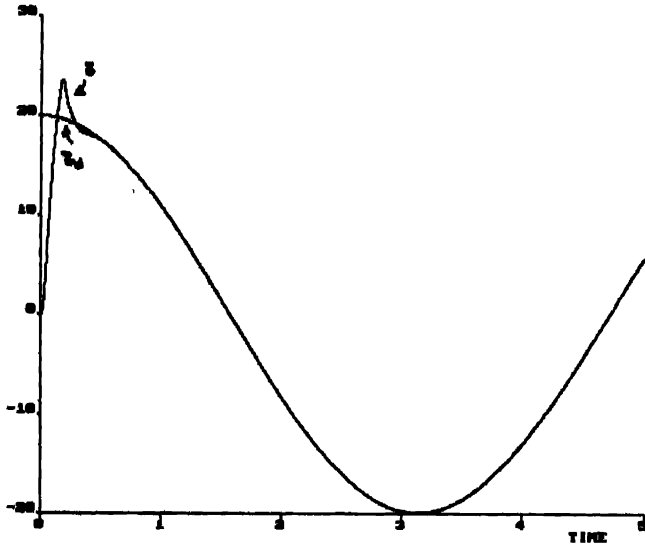


Fig. 1. Generated and desired torque.

4) The ab model is sometimes given in terms of stator currents and rotor fluxes, instead of stator and rotor currents. It is straightforward to verify that applying to (4.1) the linear transformation

$$z = \begin{bmatrix} I_2 & 0 & 0 \\ L_s I_2 & L_s I_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} x$$

and choosing $\omega_n = 0$, one obtains the model studied in several recent papers, e.g., [14] and [6].

V. SIMULATION RESULTS

The performance of the control scheme of Proposition 4.3 was investigated by simulations. The numerical values of the four-pole squirrel-cage induction motor used in [7] were chosen, that is, $R_s = 0.687 \Omega$, $R_r = 0.842 \Omega$, $L_s = 84 \text{ mH}$, $L_r = 85.2 \text{ mH}$, $L_{sr} = 81.3 \text{ mH}$, $J = 0.03 \text{ Kg}^2$, and $b = 0.01 \text{ kg}^2 \text{ s}^{-1}$. We present here simulations of a sinusoidal torque reference with load torque $\tau_L = (2.75 + 0.003\dot{q}_1^2) \text{sgn}(\dot{q}_1) + 0.15\dot{q}_1$.

The motor is initially in stand-still with zero initial conditions. The controller is initialized with the values $\lambda_{i,d}(0) = [1, 0]^T$, $\dot{q}_i(0) = [10, 5, -5, 10]^T$, and the parameters $\beta = 1$, $\epsilon = 0.5$. Fig. 1 shows the response of the generated and the desired torque, and in Fig. 2 the rotor currents and their observed values are depicted. Notice that although the observation error converges to zero in less than 0.3 s, output tracking takes longer because of the slower state error convergences, which are shown in Fig. 3. The required control effort is very smooth as shown in Figs. 4 and 5. As seen from the figures, and taking into account the fact that the simulation includes the motor start-up and zero crossing of the generated torque with a highly nonlinear load torque characteristic, the performance of the closed-loop system is quite remarkable.

VI. CONCLUDING REMARKS

We have solved in this note the output feedback torque tracking and flux regulation problem for an induction motor model given in physical $\alpha\beta$ frame. The controller is based on the "energy-shaping" methodology of [16]. It is shown that it can be obtained by "rotating-back" the control law of [17] or vice versa.

The following features of the proposed controller are worth emphasizing.

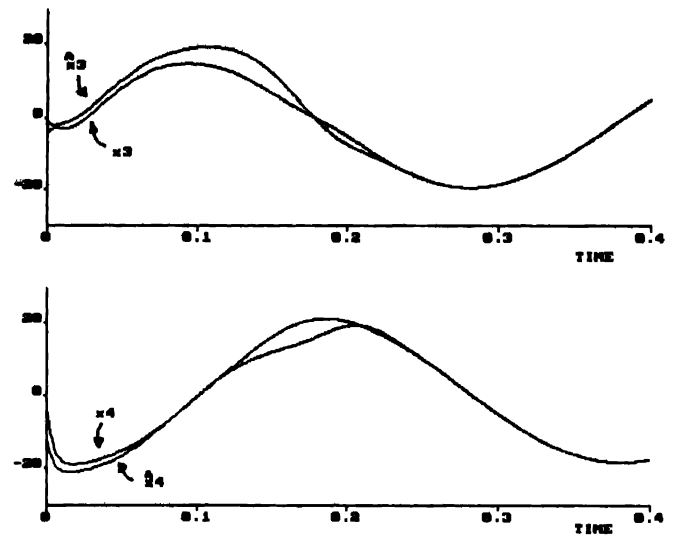


Fig. 2. Real and observed rotor currents.

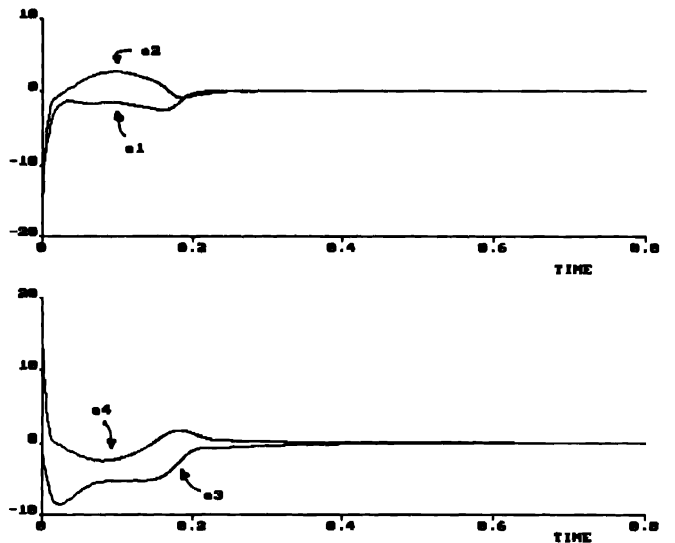


Fig. 3. Electrical state error signals.

- In contrast with [14], we *do not* require measurement of rotor variables, typically unavailable in applications.
- The control law is *always well defined* (i.e., there are no singular points), even in start-up. This constitutes an unquestionable advantage with respect to existing results, e.g., [6], where the set of singularities is determined by the design methodology, and consequently is unrelated to the physical structure of the system and very difficult to analyze.
- Tuning and commissioning are very simple. There is essentially only one design parameter ϵ in (3.10), which reflects the confidence in our rotor resistance estimate. Since this estimate is usually bad, we essentially are required to inject some high gain in the loop.
- To overcome the need of knowing the torque load, we can use a parameter estimator as done in [16] and [17]. A more challenging problem is uncertainty in the motor resistances, for which solutions are known only with full state measurement [16], [14]. It is worth pointing out that in [4] some simulations that reveal instability of a scheme including parameter adaptation and (a globally stable) state estimator have been reported.

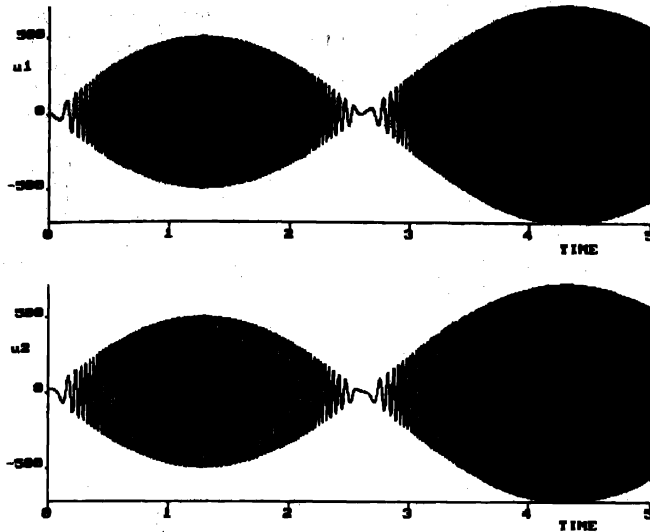


Fig. 4. Stator voltages.

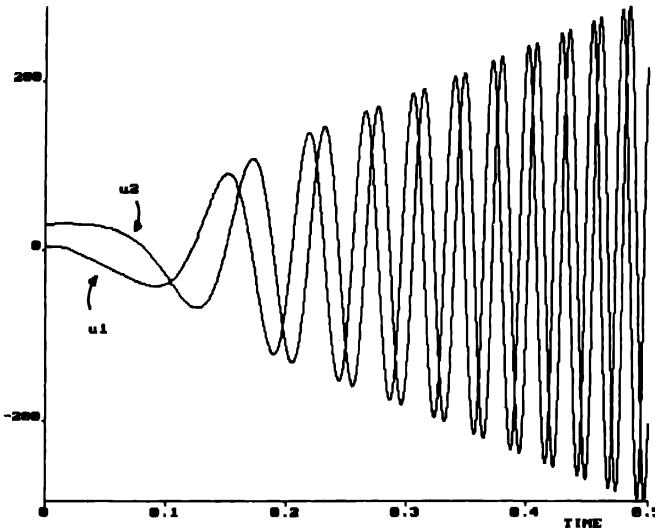


Fig. 5. A window over the stator voltages.

For speed/position control applications, a slight modification to the present controller must be made. This has been recently reported in [18].

APPENDIX PROOF OF MAIN RESULT

First, notice that the controller (3.7) is obtained from the substitution of (3.2), (3.3) into (3.5). Now, the error equation is given by

$$D(q_5)\dot{e} + C(q_5, \dot{q}_5)e + (\mathcal{R} + \mathcal{K}(q_5, \dot{q}_5))e = S(q_5, \dot{q}_5)\dot{q}_d \quad (\text{A.1})$$

where $\mathcal{K} = \text{diag}\{K_1(\dot{q}_5)I_2, 0, K_2(\dot{q}_5)\}$ and

$$S = \begin{bmatrix} 0 & L_{sr} \mathcal{T} e^{J q_5} \dot{q}_{5,d} \\ 0 & 0 \\ 0 & L_{sr} \dot{q}_{5,d}^T \mathcal{T} e^{J q_5} \end{bmatrix} \in \mathcal{R}^{5 \times 4}.$$

On the other hand, from (3.11) and (2.3), the observation error satisfies the following equation:

$$D_r(q_5)\ddot{\hat{q}}_r + (W_1(q_5)\dot{\hat{q}}_r + L(q_5, \dot{\hat{q}}_5))\dot{\hat{q}}_r + R_r \dot{\hat{q}}_r = 0. \quad (\text{A.2})$$

Consider the quadratic function $V = \frac{1}{2}e^T D(q)e + \frac{1}{2}\dot{q}_r^T \dot{q}_r$ whose time derivative, taking into account the antisymmetry of $\dot{D} - 2C$ and $\dot{D}_r - 2(W_1\dot{q}_5 + L)$, yields

$$\dot{V} = -e^T (\mathcal{R} + \mathcal{K})e + e^T S \dot{q}_r - \dot{q}_r^T R_r \dot{q}_r = -z^T Q z$$

where we have defined $z = [e^T, \dot{q}_r^T]^T$, and

$$Q = \begin{bmatrix} \mathcal{R} + \mathcal{K} & \frac{1}{2}S \\ \frac{1}{2}S^T & R_r \end{bmatrix}.$$

Checking that (3.10) ensures positive definiteness of Q , we conclude that e and \dot{q}_r are bounded and converge to zero. Internal stability is established noting that $\lambda_{r,d}$ is bounded by construction so \dot{q}_r is bounded. Boundedness of $\dot{q}_{5,d}$ follows from the fact that $\dot{q}_{5,d}$ is the output of an asymptotically stable filter with bounded inputs (3.9). $\square\square\square$

REFERENCES

- [1] C. Canudas, R. Ortega, and S. Seleme, "Robot motion control using induction motor drives," in *Proc. IEEE ICRA 93*, Atlanta, GA, May 4-6, 1993.
- [2] J. Chiasson, "Dynamic feedback linearization of the induction motor," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 1588-1594, Oct. 1993.
- [3] A. Deluca, "Design of an exact nonlinear controller for induction motors," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 1304-1307, Dec. 1989.
- [4] G. Espinosa, "Nonlinear control of induction motors," Ph.D. dissertation, UNAM, 1993.
- [5] G. Espinosa and R. Ortega, "State observers are unnecessary for induction motor control," *Syst. Contr. Lett.*, vol. 23, pp. 315-323, 1994.
- [6] I. Kanelakopoulos, P. Krein, and F. Disilvestro, "Nonlinear flux-observer-based control of induction motors," in *Proc. ACC'92*, Chicago, IL, May 1992.
- [7] D. Kim, I. Ha, and M. Ko, "Control of induction motors via feedback linearization with input-output decoupling," *Int. J. Contr.*, vol. 51, no. 4, pp. 863-883, 1990.
- [8] P. C. Krause, *Analysis of Electric Machinery*. New York: McGraw-Hill, 1986.
- [9] Z. Krzeminski, "Nonlinear control of induction motor," in *Proc. 10th IFAC World Congr.*, Munich, 1987, pp. 349-354.
- [10] H. Kwatny and H. Kim, "Variable structure regulation of partially linearizable systems," *Syst. Contr. Lett.*, vol. 15, no. 1, pp. 67-80, July 1990.
- [11] W. Leonhard, *Control of Electrical Drives*. Berlin: Springer-Verlag, 1985.
- [12] X. Liu, G. Verghese, J. Lang, and M. Onder, "Generalizing the Blondel-Park transformation of electrical machines," *IEEE Trans. Circ. Syst.*, vol. 36, pp. 1058-1067, 1989.
- [13] R. Marino, S. Peresada, and P. Tomei, "Adaptive output feedback control of current-fed induction motors," in *Proc. IFAC World Congr.*, Sydney, 1993.
- [14] R. Marino, S. Peresada, and P. Valigi, "Adaptive partial feedback linearization of induction motors," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 208-221, Feb. 1993.
- [15] J. Meisel, *Principles of Electromechanical Energy Conversion*. New York: McGraw-Hill, 1969.
- [16] R. Ortega and G. Espinosa, "Torque regulation of induction motors," *Automatica*, vol. 29, no. 3, pp. 621-633, 1993.
- [17] R. Ortega, C. Canudas, and S. Seleme, "Nonlinear control of induction motors: Torque tracking with unknown load disturbances," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 1675-1680, Nov. 1993.
- [18] R. Ortega, P. J. Nicklasson, and G. Espinosa, "On speed control of induction motors," submitted to *Automatica*.
- [19] R. Ortega and M. Spong, "Adaptive motion control of rigid robots: A tutorial," *Automatica*, vol. 25, no. 6, pp. 877-888, 1989.
- [20] M. Spong, and M. Vidyasagar, *Robot Dynamics and Control*. New York: Wiley, 1989.
- [21] M. J. Riezenman, "Electric vehicles," *IEEE Spectrum*, vol. 29, 1992.
- [22] S. Seely, *Electromechanical Energy Conversion*. New York: McGraw-Hill, 1962.
- [23] V. Utkin, "Sliding mode control design principles and applications to electric drives," *IEEE Trans. Ind. Electron.*, vol. 40, pp. 23-36, Feb. 1993.

Eigenstructure Assignment in Linear Descriptor Systems

Petr Zagalak and Vladimír Kučera

Abstract—Given a linear controllable system $E\dot{x} = Fx + Gu$, we study the limits of linear proportional feedback $u = Kx + v$ in assigning a desired finite and infinite eigenstructure of the closed-loop system $E\dot{x} = (F + GK)x + Gv$. The closed-loop system is not restricted to be regular; its rank makes part of the specifications. Necessary and sufficient conditions are established for the existence of a feedback gain K which assigns the desired eigenstructure, and a procedure is outlined to calculate one such gain.

I. INTRODUCTION

Let

$$E\dot{x} = Fx + Gu \quad (1)$$

be a linear descriptor system where E , F are $n \times n$ matrices, and G is an $n \times m$ matrix of rank m over R , the field of real numbers.

The problem of eigenstructure assignment by linear proportional feedback

$$u = Kx + v \quad (2)$$

where K is an $m \times n$ matrix over R , has had a long history. When E is nonsingular, the eigenstructure of (1) is specified by the invariant polynomials of $sE - F$, or equivalently by the position and structure of the system's finite eigenvalues. Then Rosenbrock [12] showed that a controllable system (1) with controllability indexes c_1, c_2, \dots, c_m (decreasingly ordered by magnitude) can be assigned the eigenstructure given by invariant polynomials $\psi_1(s), \psi_2(s), \dots, \psi_m(s)$ (decreasingly ordered by degree) if and only if the set of inequalities

$$\sum_{i=j}^m \deg \psi_i(s) \leq \sum_{i=j}^m c_i, \quad j = 1, 2, \dots, m \quad (3)$$

are satisfied and equality holds when $j = 1$. This famous result, commonly referred to as the fundamental theorem of state feedback, identifies the limits of feedback (2) in altering the eigenstructure of (1). Alternative proofs were given later by Dickinson [1], Flamm [2], and Kučera [5].

A more general situation arises when E is singular. Then the pencil $sE - F$ has finite as well as infinite elementary divisors, or, equivalently, the system has finite as well as infinite eigenvalues. This means that the system gives rise to exponential as well as impulsive modes. Assuming that $sE - F$ is regular, Kučera and Zagalak [6] generalized Rosenbrock's result in the situation where the desired eigenstructure is specified solely by finite elementary divisors or by invariant polynomials. The other extreme, where only infinite elementary divisors are being assigned, was considered by Zagalak and Kučera [13].

The problem of simultaneously assigning the finite and the infinite eigenstructure was solved by the same authors in [14]. The result says that a regular controllable system (1) with controllability indexes c_1, c_2, \dots, c_m (decreasingly ordered by magnitude) can be assigned

by applying feedback (2) that preserves its regularity, the eigenstructure specified by invariant polynomials $\psi_1(s), \psi_2(s), \dots, \psi_m(s)$ (decreasingly ordered by degree), and by infinite eigenvalue orders d_1, d_2, \dots, d_m (decreasingly ordered by magnitude) if and only if the set of inequalities

$$\sum_{i=j}^m \deg \psi_i(s) + d_i \leq \sum_{i=j}^m c_i, \quad j = 1, 2, \dots, m \quad (4)$$

are satisfied with equality holding when $j = 1$, and

$$d_i = 0, \quad i > m - q \quad (5)$$

where q is the number of infinite zeros of (1). The necessity of (4) was proven already by Jones, Pugh, and Hayton [3] and conditions equivalent to (4) and (5) were given, without proof, in Özçaldıran [10].

The requirement that the resulting feedback system

$$E\dot{x} = (F + GK)x + Gv \quad (6)$$

be regular is too restrictive in many situations. That is why we relax this assumption here and make the rank of $sE - (F + GK)$ a part of the specifications. This will allow for desired eigenstructures specified by a number of invariant polynomials less than n .

II. BASIC CONCEPTS

The system (1) is said to be controllable if $\text{rank } [sE - F \ G] = n$ for all complex s , finite and infinite. For details, see Lewis [8] and Verghese, Lévy, and Kailath [12], and the references given therein.

Let $N(s)$, $D(s)$ be, respectively, $n \times m$, $m \times m$ matrices over $R[s]$, the ring of polynomials in s with coefficients in R , such that

$$[sE - F \ -G] \begin{bmatrix} N(s) \\ D(s) \end{bmatrix} = 0.$$

Then $N(s)$, $D(s)$ are said to form a (right) normal external description of (1) if

i)

$$\begin{bmatrix} N(s) \\ D(s) \end{bmatrix}$$

is a decreasingly column-degree ordered, minimal polynomial basis of $\text{Ker } [sE - F \ -G]$; and

ii) $N(s)$ is a minimal polynomial basis of $A(sE - F)$, where A is a maximal annihilator of G .

See Malabre, Kučera, and Zagalak [9] for details.

Then the controllability indexes of (1) are the column degrees

$$c_i = \deg_i \begin{bmatrix} N(s) \\ D(s) \end{bmatrix}, \quad i = 1, 2, \dots, m$$

of any normal external description of (1). When $\deg_i D(s) > \deg_i N(s)$, we call c_i a proper controllability index of (1).

III. PROBLEM FORMULATION

Let (1) be a controllable system with controllability indexes $c_1 \geq c_2 \geq \dots \geq c_m$, and let q be the number of the proper controllability indexes. We are given r monic polynomials $\psi_1(s), \psi_2(s), \dots, \psi_r(s)$ in $R[s]$ such that $\psi_{i+1}(s)$ divides $\psi_i(s)$, $i = 1, 2, \dots, r-1$, and

Manuscript received February 24, 1992; revised May 2, 1994. This work was supported by the Czechoslovak Academy of Sciences under Grant 27 501. The original version of this paper was presented at the 1st European Control Conference, Grenoble, France, July 2-5, 1991.

The authors are with the Institute of Information Theory and Automation, Academy of Sciences, 182 08 Prague, Czech Republic.

IEEE Log Number 9407029.

positive integers $d_1 \geq d_2 \geq \dots \geq d_l$. These polynomials and integers are not arbitrary, they are assumed to satisfy

$$\sum_{i=1}^l \deg \epsilon_i(s) + \sum_{i=1}^l d_i = \sum_{i=1}^+ \epsilon_i^* \quad (7)$$

for some selection $\epsilon_1^* \geq \epsilon_2^* \geq \dots \geq \epsilon_{l+m-n}^*$ of $l+m-n$ controllability indexes of (1) that includes all the proper controllability indexes.

The problem of eigenstructure assignment is then as follows. First establish necessary and sufficient conditions under which a feedback gain K exists in (2) such that $\epsilon_i(s)$'s are the invariant polynomials and d_i 's are the infinite eigenvalue orders of the closed loop system (6). Second if the conditions are verified give a procedure to calculate one such K .

IV. FUNDAMENTAL THEOREM

The following theorem solves the existence part of the eigenstructure assignment problem.

Theorem The eigenstructure assignment problem has a solution if and only if

$$n - m + p + q \leq l \leq n \quad (8)$$

and

$$\sum_{i=1}^l \deg \epsilon_i(s) + d_i \leq \sum_{i=1}^+ \epsilon_i^* \quad i = 1, 2, \dots, l \quad (9)$$

where by convention $d_i = 0$ for $i < p$ and $\epsilon_i^* = 0$ for $i > l+m-n$.

Proof (Necessity) We suppose there exists a feedback gain K such that the eigenstructure of (6) is given by the lists $\epsilon_1(s), \dots, \epsilon_l(s)$ and d_1, \dots, d_l that satisfy (7) for some $\epsilon_1^* \geq \epsilon_2^* \geq \dots \geq \epsilon_{l+m-n}^*$. We wish to prove (8) and (9).

To handle the finite and the infinite eigenstructure in a uniform way we shall apply the conformal mapping

$$u = \frac{s}{s-a} \quad (10)$$

where $a \neq 0$ is a complex number that is not an eigenvalue of (6).

Let $\Lambda(s) = D(s)$ be a normal external description of (1). Then

$$[sI - F - G(K - G)] \begin{bmatrix} \Lambda(s) \\ D(s) - K\Lambda(s) \end{bmatrix} = 0 \quad (11)$$

so that $\Lambda(s) = D(s) - K\Lambda(s)$ is a normal external description of the closed loop system (6). Since controllability is invariant under the action of feedback (2) the systems (1) and (6) are controllable with the same controllability indexes.

We define

$$\begin{bmatrix} B(u) \\ C(u) \end{bmatrix} = \begin{bmatrix} \Lambda\left(\frac{u}{u-1}\right) \\ D\left(\frac{u}{u-1}\right) - K\Lambda\left(\frac{u}{u-1}\right) \end{bmatrix} \quad \text{diag}[(u-1)^{-1}, \dots, (u-1)^{-1}] \quad (12)$$

This is a polynomial matrix over $R[u]$ which is irreducible and column proper with column degrees $\epsilon_1, \epsilon_2, \dots, \epsilon_l$. Obviously, $C(u)$ also has column degrees $\epsilon_1, \epsilon_2, \dots, \epsilon_l$. We suppose without loss of generality that the matrix F has been brought to the form

$$E = \begin{bmatrix} E_1 \\ 0 \end{bmatrix} \quad (13)$$

with F_1 of full row rank. Say f . We define

$$P(u) = -Q(u) = \text{diag} \left[\underbrace{u-1, \dots, u-1}_f, 1, \dots, 1 \right] \left[\frac{au}{u-1} F - (F + G)K - G \right]$$

This is again a polynomial matrix over $R[u]$ which is irreducible and row proper.

It follows from (11) that

$$[P(u) - Q(u)] \begin{bmatrix} B(u) \\ C(u) \end{bmatrix} = 0$$

Lemma 1 of the Appendix implies that the rank deficiency of $P(u)$ and $Q(u)$ are the same, namely

$$\text{rank } C(u) = l + m - n$$

since $\text{rank } P(u) = l$ by assumption. Further the invariant polynomials of $B(u)$ equal those of $Q(u)$, namely

$$\epsilon_1(u) = 1, \dots, \epsilon_f(u) = 1$$

$$\epsilon_{f+1}(u) = u-1, \dots, \epsilon_{f+m-n}(u) = u-1$$

$$\epsilon_{f+m-n+1}(u) = 1, \dots, \epsilon_l(u) = 1$$

and the nonunit invariant polynomials of $C(u)$ are equal to those of $P(u)$.

$$\omega_1(u) = \bar{\epsilon}_1(u)(u-1)^{l_1}, \dots, \omega_f(u) = \bar{\epsilon}_f(u)(u-1)^{l_f}$$

$$\omega_{f+1}(u) = \bar{\epsilon}_{f+1}(u), \dots, \omega_{f+m-n}(u) = \bar{\epsilon}_{f+m-n}(u) \quad (14)$$

where

$$\bar{\epsilon}_i(u) = (u-1)^{l_i - \epsilon_i} \epsilon_i \left(\frac{-au}{u-1} \right) \quad (15)$$

Since (12) is irreducible $\epsilon_i(u)$ is prime to $\omega_j(u)$ $i = 1, 2, \dots, f + m - n$. Hence $p \leq f + m - n - q$. Since $l \leq n$ we have (8).

Now suppose without loss of generality that it is the first $f + m - n$ columns of $C(u)$ that are $R(u)$ linearly independent. Let $\bar{C}(u)$ be the submatrix of $C(u)$ formed by these columns. Then by assumption (7) $\bar{C}(u)$ is column proper with column degrees $\epsilon_1^* \geq \epsilon_2^* \geq \dots \geq \epsilon_{f+m-n}^*$. Then the product $\omega_j(u) = \omega_{f+m-n+1}(u)$ is the greatest common divisor of all minors of $\bar{C}(u)$ of order $f + m - n - j + 1$ which implies

$$\sum_{i=1}^+ \deg \omega_i(u) = \sum_{i=1}^+ \deg \bar{\epsilon}_i(u) + d \leq \sum_{i=1}^+ \epsilon_i^* \quad i = 1, \dots, f + m - n \quad (16)$$

where $d = 0$ for $i > p$ has been defined for convenience of summation.

Lemma 1 implies that $\bar{\epsilon}_i(u) = 1$ for $i = f + m - n + 1, \dots, l$. Adopting the convention that $\epsilon_i^* = 0$ for $i > f + m - n$ we can extend the upper summation limit in (16) to l . Finally since $\deg \bar{\epsilon}_i(u) = \deg \epsilon_i(s)$ $i = 1, 2, \dots, l$ the inequalities (9) are verified.

(Sufficiency) We consider a list of monic polynomials $\epsilon_1(s), \epsilon_2(s), \dots, \epsilon_l(s)$ in $R[s]$, a list of positive integers d_1, d_2, \dots, d_l and a subset $\epsilon_1^*, \epsilon_2^*, \dots, \epsilon_{l+m-n}^*$ of the controlability indexes $\epsilon_1, \epsilon_2, \dots, \epsilon_l$ that satisfy (7), (8) and (9). We are going to construct a matrix K over R such that the eigenstructure of (6) will be given by these two lists.

To this end we apply to $\epsilon_i(s)$ the inverse transformation of (10) to get the polynomials $\bar{\epsilon}_i(u)$ defined in (15). Clearly $\deg \bar{\epsilon}_i(u) = \deg \epsilon_i(s)$ $i = 1, 2, \dots, l$. It follows from (9) for $j = f + m - n + 1$ that $\bar{\epsilon}_i(u) = \epsilon_i(s) = 1$ for $i > f + m - n$.

Let $N(s)$, $D(s)$ be a normal external description of (1). Applying the same transformation to

$$\begin{bmatrix} N(s) \\ D(s) \end{bmatrix}$$

and $[sE - F - G]$, where E assumes the form (13), we obtain matrices $A(w)$, $B(w)$ that satisfy

$$\text{diag}[\underbrace{w-1, \dots, w-1}_f, 1, \dots, 1] \left[\frac{aw}{w-1} E - F - G \right] \cdot \begin{bmatrix} B(w) \\ A(w) \end{bmatrix} = 0. \quad (17)$$

As in (12),

$$\begin{bmatrix} B(w) \\ A(w) \end{bmatrix}$$

is a polynomial matrix over $R[w]$ that is irreducible and column proper with column degrees c_1, c_2, \dots, c_m .

Now we construct the matrix

$$\tilde{C}(w) = \text{diag}[\phi_1(w), \dots, \phi_{r+m-n}(w)]$$

where $\phi_i(w)$ are the polynomials defined in (14) and (15). In view of (7) and (9), we can apply to $\tilde{C}(w)$ Lemma 2 of the Appendix. There results a matrix $\hat{C}(w)$ having the same invariant polynomials $\phi_1(w), \dots, \phi_{r+m-n}(w)$, and that is column proper with column degrees $c_1^* \geq c_2^* \geq \dots \geq c_{r+m-n}^*$. We further define the $m \times m$ matrix

$$\hat{C}(w) = \begin{bmatrix} \hat{C}(w) & 0 \\ 0 & 0 \end{bmatrix}.$$

Let S be a permutation matrix such that the first $r+m-n$ columns of $B(w)S$ have degrees $c_1^*, c_2^*, \dots, c_{r+m-n}^*$. Consider the matrix

$$\begin{bmatrix} B(w)S \\ \hat{C}(w) \end{bmatrix}. \quad (18)$$

If (18) is irreducible, we put $C(w) = \hat{C}(w)S^{-1}$. If it is not, there must be a zero at $w = 1$ common to $B(w)S$ and $\hat{C}(w)$. This implies a zero column among the first p columns of (18). As (8) holds and the last $r+m-n-p$ columns of $\hat{C}(w)$ are linearly independent, there exists a nonsingular and constant matrix T such that

$$\begin{bmatrix} B(w)S \\ \hat{C}(w)T \end{bmatrix}$$

is irreducible. Then we put $C(w) = \hat{C}(w)TS^{-1}$. In either case, the matrix

$$\begin{bmatrix} B(w) \\ C(w) \end{bmatrix} \quad (19)$$

is irreducible and column proper with column degrees c_1, c_2, \dots, c_m . Applying the transformation (10) to (19), we obtain the matrix

$$\begin{bmatrix} N(s) \\ H(s) \end{bmatrix} = \begin{bmatrix} B \left(\frac{s}{s-a} \right) \\ C \left(\frac{s}{s-a} \right) \end{bmatrix} \text{diag}[(s-a)^{-1}, \dots, (s-a)^{-m}].$$

This matrix is irreducible, column proper, and its column degrees are equal to c_1, c_2, \dots, c_m .

Let us now consider the equation

$$XD(s) + YN(s) = H(s). \quad (20)$$

Since the matrices

$$\begin{bmatrix} N(s) \\ D(s) \end{bmatrix}, \quad \begin{bmatrix} N(s) \\ H(s) \end{bmatrix}$$

are both irreducible and column proper with column degrees c_1, c_2, \dots, c_m and $N(s)$, $D(s)$ is a normal external description,

the rows of these matrices span the same R -linear spaces. Then, by Lemma 3 of the Appendix, equation (20) has a constant solution pair X, Y such that X is nonsingular. Thus, $K = -X^{-1}Y$ qualifies as a feedback gain that assigns the desired eigenstructure to $sE - (F + GK)$. This immediately follows from Lemma 1.

V. CONSTRUCTION

We summarize the major steps of the sufficiency part of the proof of the Theorem, which provide a construction for K .

a) Given E, F, G with E having the form (13), calculate polynomial matrices $A(w), B(w)$ such that (17) holds where $B(w)$ and

$$\begin{bmatrix} B(w) \\ A(w) \end{bmatrix}$$

are irreducible and column proper, and

$$\begin{bmatrix} B(w) \\ A(w) \end{bmatrix}$$

is decreasingly column-degree ordered.

b) Read out c_1, c_2, \dots, c_m , the column degrees of

$$\begin{bmatrix} B(w) \\ A(w) \end{bmatrix}$$

and identify the subset of q proper controllability indexes.

c) Choose $\psi_1(s), \psi_2(s), \dots, \psi_r(s)$ and d_1, d_2, \dots, d_p and $c_1^*, c_2^*, \dots, c_{r+m-n}^*$ that satisfy (7) and check for the existence of K using (8) and (9).

d) Construct an $m \times m$ polynomial matrix $C'(w)$ having the invariant polynomials $\phi_i(w)$ defined in (14) and (15), and such that the composite matrix

$$\begin{bmatrix} B(w) \\ A(w) \end{bmatrix}$$

is irreducible and column proper with column degrees c_1, c_2, \dots, c_m .

e) Find a constant solution pair X, Y of the equation

$$XA(w) + YB(w) = C'(w)$$

such that X is nonsingular.

f) Put $K = -X^{-1}Y$.

Of course, there may be feedback gains other than K that solve the problem.

VI. CONCLUSIONS

The limits of proportional state feedback (2) in altering the eigenstructure of the linear controllable system (1) have been established in Theorem 1, under the assumption (7) that the total number of the dynamical modes desired is equal to a sum of $r+m-n$ controllability indexes. These limits can be summarized as follows.

- The eigenvalues can be placed at any position.
- The measure of regularity $r = \text{rank}(sE - F - GK)$ is bounded by (8).
- At most, $r+m-n$ cyclic chains can be associated with each finite eigenvalue.
- At most, $r+m-n-q$ cyclic chains can be associated with the infinite eigenvalue, because q is the number of infinite zero chains.
- The sizes of the cyclic chains are limited from below by (9).
- When E is nonsingular, all controllability indexes are proper, $q = m$, and no infinite eigenvalue can be assigned.

The controllability assumption can be relaxed. A generalization of the theorem to partially controllable systems is under study. We also conjecture that (7) is a necessary restriction on the eigenstructure desired whenever $r = n$.

APPENDIX

Lemma 1 [15] Let $P(s)$, $Q(s)$ and $B(s)$, $C(s)$ be polynomial matrices over $R[s]$ of respective size $n \times n$, $n \times m$ and $n \times m$, $m \times m$ such that

$$[P(s) \quad -Q(s)] \begin{bmatrix} B(s) \\ C(s) \end{bmatrix} = 0$$

here

$$[P(s) \quad -Q(s)] \text{ and } \begin{bmatrix} B(s) \\ C(s) \end{bmatrix}$$

are irreducible, and $\text{rank } Q(s) = m$. Then,

- the matrices $C(s)$ and $P(s)$ have the same rank deficiency
- the matrices $B(s)$ and $Q(s)$ have the same invariant polynomials
- the matrices $C(s)$ and $P(s)$ have the same nonunit invariant polynomials

Lemma 2 [15] Let $\bar{C}(s)$ be a column proper polynomial $l \times l$ matrix over $R[s]$ and let $\alpha_1(s) \alpha_2(s) \dots \alpha_k(s)$ be its invariant polynomials arranged so that $\alpha_{j+1}(s)$ divides $\alpha_j(s)$, $j = 1, 2, \dots, l-1$. Let $\epsilon_1 \geq \epsilon_2 \geq \dots \geq \epsilon_k$ be nonnegative integers such that

$$\sum_{i=1}^l \deg \alpha_i(s) = \sum_{i=1}^k \epsilon_i, \quad j = 1, 2, \dots, l$$

with equality holding when $j = 1$.

Then there exist unimodular matrices $U_1(s)$ and $U_2(s)$ such that the matrix

$$\bar{C}(s) = U_1(s)C(s)U_2(s)$$

is column proper with column degrees $\epsilon_1, \epsilon_2, \dots, \epsilon_k$.

Lemma 3 [7] Let $D(s)$, $N(s)$ and $H(s)$ be respectively $m \times m$, $n \times m$ and $m \times m$ polynomial matrices over $R[s]$. Then the equation

$$N(s)D(s) + Y(s)N(s) = H(s)$$

has a solution pair N, Y over R such that N is nonsingular if and only if the rows of the matrices

$$\begin{bmatrix} N(s) \\ D(s) \end{bmatrix} \quad \begin{bmatrix} N(s) \\ H(s) \end{bmatrix}$$

span the same R linear spaces.

REFERENCES

- [1] B. W. Dickinson, On the fundamental theorem of linear state variable feedback, *IEEE Trans Automat Contr*, vol. AC-19, pp. 577-579, 1974.
- [2] D. S. Flamm, A new proof of Rosenbrock's theorem on pole assignment, *IEEE Trans Automat Contr*, vol. AC-25, pp. 1128-1133, 1980.
- [3] F. R. I. Jones, A. C. Pugh and G. F. Hayton, Necessary conditions for the general pole placement problem via constant output feedback, *Int J Contr*, vol. 51, pp. 771-784, 1990.
- [4] T. Kailath, *Linear Systems*, Englewood Cliffs, NJ: Prentice Hall, 1980.
- [5] V. Kučera, Assigning the invariant factors by feedback, *Kybernetika*, vol. 17, pp. 118-127, 1981.
- [6] V. Kučera and P. Zagalak, Fundamental theorem of state feedback for singular systems, *Automatica*, vol. 24, pp. 653-658, 1988.
- [7] —, Constant solutions of polynomial equations, *Int J Contr*, vol. 53, pp. 495-502, 1991.
- [8] F. I. Lewis, A survey of linear singular systems, *J Circuits Syst Signal Proc*, Special Issue on Singular Systems, vol. 5, pp. 3-36, 1986.

- [9] M. Mulabre, V. Kučera and P. Zagalak, Reachability indices for linear descriptor systems, *Syst. C*, pp. 119-123, 1990.
- [10] K. Ozaldiran, Fundamental theorem of linear state singular systems, in *Proc. 29th Conf. Dec. Comm. H. J. J. J.*
- [11] H. H. Rosenbrock, *State Space and Multivariable Theory*, Wiley, 1970.
- [12] G. C. Verghese, B. C. Levy and T. Kailath, A generalization for singular systems, *IEEE Trans Automat Contr*, vol. AC-26, pp. 811-831, 1981.
- [13] P. Zagalak and V. Kučera, Fundamental theorem of state feedback: The case of infinite poles, *Kybernetika*, vol. 27, pp. 1-11, 1991.
- [14] —, Fundamental theorem of proportional state feedback for descriptor systems, *Kybernetika*, vol. 28, pp. 81-89, 1992.
- [15] —, Eigenstructure assignment in linear descriptor systems, in *Proc. 1st Euro Contr. Conf.*, Grenoble, 1991, pp. 409-414.

Robust Regulation in the Presence of Norm-Bounded Uncertainty

J. Abedor, K. Nagpal, P. P. Khargonekar and K. Poolla

Abstract—We consider robust regulation (against steps and sinusoids) in the presence of unstructured uncertainty. The unstructured uncertainty is norm bounded by a constant that is given *a priori*. This problem is equivalent to a certain multiobjective problem where one objective is robust regulation and the other is the standard objective of \mathcal{H}_∞ suboptimal control. It is shown that a solution to this problem exists if and only if the standard \mathcal{H}_∞ problem admits a solution and certain matrix inequalities are satisfied. These solvability conditions are readily computable. Controller synthesis is also addressed.

1. INTRODUCTION

Before discussing the problem that is the focus of this note, we briefly review the standard problem of robust regulation. This problem will be defined with respect to Fig. 1. Here P is the plant and K is the controller to be designed. Both are finite dimensional linear time invariant systems, as are all systems throughout this note. In the figure, u is the exogenous input, y is the controlled output, v is the control input and q is the measurement. The signal q is assumed to be measured. Without loss of generality we assume, in fact, that q consists of the first l channels of y .

Suppose now that we are interested in regulating against sinusoids with frequencies $\omega_1, \omega_2, \dots, \omega_N$. Let \mathcal{F} denote the vector space of signals consisting of linear combinations of these sinusoids (\mathcal{F} will include the class of constant signals if one of these frequencies is zero). The robust regulation synthesis problem is the problem of designing an internally stabilizing controller such that the closed

Manuscript received March 5, 1993; revised February 2, 1994. This work was supported in part by National Science Foundation Grants ECS 9001371 and ECS 8957461, Air Force Office of Scientific Research Grant AFOSR 90-0053, Army Research Office Grant DAAH04-93-C-0012, and by Rockwell International.

J. Abedor is with Integrated Systems, Inc., Santa Clara, CA 95054 USA. K. Nagpal and K. Poolla are with the Department of Mechanical Engineering, University of California, Berkeley, CA 94720 USA.

P. P. Khargonekar is with the Department of Electrical Engineering, University of Michigan, Ann Arbor, MI 48109 USA.

IEEE Log Number 9407031.

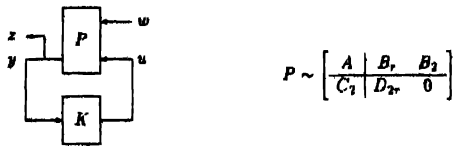


Fig. 1. Setup for the robust regulation problem.

loop system asymptotically rejects \mathcal{F} from w to z , and does so locally robustly. The problem, in other words, is to design K so that

- 1) K internally stabilizes P .
- 2) $T_{zw}(j\omega_k) = 0$ for $k = 1, \dots, N$, and
- 3) property 2) holds for all plants in some neighborhood of P in the graph topology.

In part 2), T_{zw} is the closed-loop transfer function from w to z . The requirement that this transfer function evaluate to zero at frequencies $\omega_1, \dots, \omega_N$ is equivalent to the requirement that \mathcal{F} be asymptotically rejected from w to z . Part 3) is the local robustness requirement.

Now consider Fig. 2. In this figure the plant P_Δ is formed from the interconnection of the nominal plant P (augmented with the additional input w_1 and the additional output z_1) and the unstructured norm-bounded uncertainty block Δ . Again K is the controller to be designed, u is the control input, and y is the measurement. The exogenous input is w_2 , and z_2 is the controlled output, which again is assumed to be measured. Without loss of generality, we again assume that z_2 consists of the first l channels of y .

The following controller synthesis problem is addressed in this paper: given a real number $\gamma > 0$, design a controller K such that for all stable Δ with $\|\Delta\|_\infty \leq 1/\gamma$,

- 1) K internally stabilizes P_Δ .
- 2) $T_{z_2 w_2}(j\omega_k) = 0$ for $k = 1, \dots, N$, and
- 3) property 2) holds for all plants in some neighborhood of P_Δ in the graph topology.

This problem will be called the problem of robust regulation in the presence of norm-bounded uncertainty (RRNBU). It differs fundamentally from the standard problem of robust regulation in that we are requiring the controller to solve the robust regulation problem for an *a priori* given (possibly large) plant uncertainty set. To emphasize this point, we note that any controller that solves the robust regulation problem for the nominal plant P will, by virtue of local robustness, solve the RRNBU problem for small enough $1/\gamma$. However there are no guarantees on the size of $1/\gamma$. The amount of unstructured uncertainty that a particular robustly regulating controller can tolerate may be arbitrarily small. Given a value γ , the results of this note enable one to efficiently check whether the corresponding RRNBU problem is solvable, and in the event that it is, construct a controller that solves the problem. Results are given for both the state-feedback and output-feedback cases.

It is shown in Section III-A that the RRNBU problem is equivalent to a certain multiobjective problem that will be discussed in the context of Fig. 3. The multiobjective problem is to design a single controller that solves both the robust regulation problem (from w_2 to z_2 , with $w_1 \equiv 0$) and the \mathcal{H}_∞ suboptimal control problem (from w_1 to z_1 , with $w_2 \equiv 0$). Precisely: given a real number $\gamma > 0$, the problem is to design a controller K such that

- 1) K internally stabilizes P .
- 2) $T_{z_2 w_2}(j\omega_k) = 0$ for $k = 1, \dots, N$.
- 3) property 2) holds for all plants in some neighborhood of P in the graph topology, and
- 4) $\|T_{z_1 w_1}\|_\infty < \gamma$.

Objectives 1) through 3) constitute the standard problem of robust regulation as defined previously. Objective 4) is the usual \mathcal{H}_∞ norm

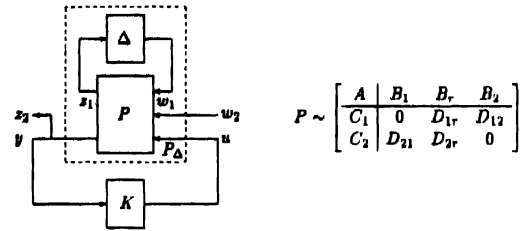


Fig. 2. Setup for the problem of robust regulation in the presence of norm-bounded uncertainty.

requirement. This multiobjective problem will be called the problem of robust regulation with an \mathcal{H}_∞ constraint (RR \mathcal{H}_∞ C).

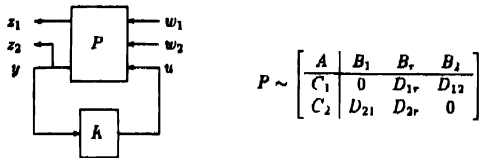
The robust regulation objective is a special case of the robust regulator problem with internal stability which was investigated extensively in the 1970s using a variety of approaches. Asymptotic tracking and asymptotic disturbance rejection are special cases of this problem. This was one of the central problems to be solved using geometric control theory. See, for example, [12], [22], [23], and the references therein. Davison and his coworkers used a matrix approach to solve this problem in [6]–[8]. The robust regulator problem with internal stability has also been investigated in the frequency domain [3], [5], [11]. One of the main results of this research is the well-known “internal model principle.” The literature on this problem is extensive and the interested reader should consult these and related references for further details.

The second objective of the RR \mathcal{H}_∞ C multiobjective problem is the standard problem of \mathcal{H}_∞ control theory initiated by Zames [25]. This problem has also been extensively investigated using both state-space and frequency-domain techniques. The book by Francis [10] contains an account of the early frequency-domain/operator-theoretic approach to this problem as well as a large bibliography. For the state-space approach, see [9], [20], and the references therein.

Problems similar to the RR \mathcal{H}_∞ C problem have been considered by Vidyasagar [21], Sugie and Vidyasagar [19], Hara and Sugie [13], [18], and Wu and Mansour [24]. In all of these notes, a frequency-domain interpolation-theoretic approach is taken. One possible shortcoming of this approach is that the resulting controllers can be of unnecessarily high McMillan degree. In an earlier conference paper [2], we gave a solution to the RR \mathcal{H}_∞ C in the special case that only steps were to be regulated against.

In this note, we take a state-space approach to the RR \mathcal{H}_∞ C problem. One merit of our approach is that the solvability conditions are given directly in terms of plant data and the solutions to at most two algebraic Riccati equations. It is shown that the RRNBU or the RR \mathcal{H}_∞ C problem is solvable if and only if the suboptimal \mathcal{H}_∞ problem is solvable and N matrix inequalities hold, where N is the number of frequencies to be regulated against. The matrices are all of size l , where l is the dimension of z_2 , i.e., the number of (scalar) signals to be regulated.

As is well known, the robust regulation objective is equivalent to requiring that the controller contain an appropriate “internal model” (see, for example, [7], [12]). Thus an obvious approach to the RR \mathcal{H}_∞ C problem is to incorporate this internal model into the generalized plant P and then solve a standard \mathcal{H}_∞ control problem. However, a significant difficulty arises—and this is why the problem is interesting—the unstable modes of the internal model are unobservable at z_1 . Since these unstable modes are on the imaginary axis, this difficulty cannot be easily circumvented. One approach is to use Riccati inequality methods as in [14]–[16]. Our results are closely related to and can be derived from the recent work of Scherer [16]; however, the control theoretic problem contexts are

Fig. 3 Setup for the problem of robust regulation with an \mathcal{H}_∞ constraint

quite different. The importance of the asymptotic regulation problem in control theory is well established, and therefore it is important to derive an explicit solution to the RRNBUC problem.

The remainder of this note is organized as follows. Notation is set in the next section, results (along with proofs) are presented in Section III, and some background material on robust regulation is collected together in the appendix.

II. NOTATION

The set of all real numbers is \mathcal{R} , the set of all imaginary numbers is $j\mathcal{R}$, and the set of all complex numbers is \mathcal{C} . For a matrix M , M' is its transpose, M^* is its conjugate transpose, $\text{spec}(M)$ is its set of eigenvalues, and $\rho(M)$ is its spectral radius. The size of a matrix will be indicated by a subscript if it is not clear from the context. We abbreviate the finite dimensional linear time invariant system $\dot{x} = Ax + Bu$, $y = Cx + Du$ by

$$I_f \sim \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right]$$

where I_f is the transfer function mapping input u to output y . Finally, a controller is said to be admissible for a plant if the resulting closed-loop system is both well posed and internally stable.

III. MAIN RESULTS

In the first part of this section we show that the RRNBUC problem and the RRH ∞ C problem are equivalent. Then, making use of this equivalence, we state and prove the main results.

A. Equivalence of RRNBUC and RRH ∞ C Problems

Theorem 3.1 Let K be any finite dimensional linear time invariant controller. Then K solves the RRNBUC problem for P if and only if K solves the RRH ∞ C problem for P .

Proof (\Rightarrow) Let K solve the RRH ∞ C problem. Then K is admissible for P and renders $\|T_{11}\|_\infty < \infty$. Thus, by the small gain theorem, the closed-loop system of Fig. 2 is internally stable for all stable Δ with $\|\Delta\|_\infty < 1/\|T_{11}\|_\infty$. By Corollary A.3, then, K solves the robust regulation problem for all stable Δ with $\|\Delta\|_\infty \leq 1/\|T_{11}\|_\infty$. Hence, K solves the RRNBUC problem.

(\Leftarrow) Suppose K solves the RRNBUC problem for P . Then, in particular, the closed-loop system of Fig. 2 is internally stable for all stable Δ such that $\|\Delta\|_\infty \leq 1/\|T_{11}\|_\infty$, which implies, by the necessity part of the small gain theorem, that $\|T_{11}\|_\infty < \infty$. Moreover, because K solves the robust regulation problem when $\Delta = 0$, K solves the robust regulation problem for P . Thus, K solves the RRH ∞ C problem.

Solution: State Feedback Case

In this section, the RRH ∞ C problem, hence the RRNBUC problem, is solved in the case where the entire state is available for feedback. We make the following standard assumptions on P :

- SF1 $C_1 = I$ and $D_{21} = 0$ (state feedback)
- SF2 (A, B_2) is stabilizable
- SF3 (C_1, A, B_1) has no uncontrollable/unobservable modes on the imaginary axis
- SF4 $D_{12}'C_1 = 0$ and $D_{11}'D_{11} = I$

The next definition enables Theorem 3.3 to be stated concisely.

Definition 3.2 (internal model matrices) A and B are internal model matrices associated with the robust regulation problem determined by $\omega_1, \dots, \omega_N$ if these matrices satisfy

- 1) $\text{spec}(A) = \{\pm j\omega_1, \dots, \pm j\omega_N\}$
- 2) every eigenvalue of A has multiplicity l
- 3) A is diagonalizable, and
- 4) (A, B) is controllable

A discussion of the role that these matrices play in robust regulation can be found in Appendix A. The state feedback result follows.

Theorem 3.3 Let the plant P of Fig. 3 satisfy the standard assumptions SF1–SF4. Then the following are equivalent:

- i) There exists a controller admissible for P that solves the robust regulation problem from u to z_1 at the frequencies $\omega_1, \dots, \omega_N$ while also making $\|T_{11}\|_\infty < \infty$ (the RRH ∞ C problem)
- ii) There exists a controller admissible for P that renders $\|T_{11}\|_\infty < \infty$ and

$$T_k B B' T_k' \preceq T_k B_1 B_1' T_k' \quad (1)$$

for all $l = 1, \dots, N$. Here $I_l = [I_l 0](j\omega_l I - A)^{-1} A = A + (B_1 B_1' - B B')\lambda$ and λ is the positive semidefinite stabilizing solution to

$$\lambda \lambda + \lambda(A + \lambda(B_1 B_1' - B B'))\lambda + C_1 C_1' = 0 \quad (2)$$

Moreover, if either (hence both) of these conditions hold, then a controller K that solves the RRH ∞ C problem for P is given by

$$K \sim \left[\begin{array}{c|c} -\frac{1}{B' I W^{-1}} & -\frac{B[I_l 0]}{B' \lambda - \frac{B[I_l 0]}{B' I W^{-1}} I} \end{array} \right] \quad (3)$$

Here I is the unique solution to $I(A - A) = B[I_l 0]$, the internal model matrices A and B satisfy 1)–4) of Definition 3.2, and W is a positive definite matrix that satisfies the Lyapunov inequality

$$AW + W(A + \lambda(B_1 B_1' - B B')) + I B B' I' = 0 \quad (4)$$

Before proving this theorem, we address the problem of constructing the controller given by (3). To synthesize this controller, one must solve a Riccati equation for λ , a linear matrix equation for I , and a linear matrix inequality (LMI) for W (computing λ and I is straightforward. (Note that I always exists since 1 and A have no eigenvalues in common.) In the simple but important case where the only frequency to be regulated against is $\omega_1 = 0$, computing W is very easy. In this case, $A = 0$, so any positive definite matrix is an acceptable W . The general case is also tractable. LMIs are finite-dimensional convex feasibility problems and thus can be readily computed [4]. We emphasize that if i) or ii) of the theorem statement hold, then we are assured that a solution to this LMI exists.

Now we turn our attention to the proof of Theorem 3.3.

Proof (i \Rightarrow ii) Suppose there exists a controller K that solves the RRH ∞ C problem for P . We first show that, as a consequence, K solves the RRH ∞ C problem for the plant

$$P_{\text{aug}} \sim \left[\begin{array}{c|c} A + B B' \lambda & B_1 - B \\ \hline B_1' \lambda & 0 \\ I & 0 \end{array} \right]$$

Because K solves the \mathcal{H}_∞ problem for P , there exists a solution λ to the algebraic Riccati equation (2). This is part of the full information

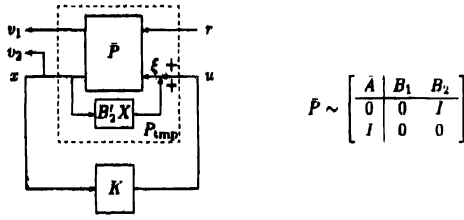


Fig. 4. The closed-loop system consisting of plant P_{imp} and controller K . P_{imp} consists of stable system \bar{P} with state-feedback $B_2'X$.

result in [9, p. 836]. Therefore by the Redheffer lemma ([9, Lemma 9]), K also solves the \mathcal{H}_∞ problem for

$$\left[\begin{array}{c|cc} A + \gamma^{-2} B_1 B_1' X & B_1 & B_2 \\ \hline B_2' X & 0 & I \\ I & 0 & 0 \end{array} \right]$$

which is precisely the plant P_{imp} . Moreover, because K solves the robust regulation problem for P , K also solves the robust regulation problem for any plant for which it is an admissible controller (because it has the right internal model). This is just a restatement of Corollary A.3. In particular, K solves the robust regulation problem for P_{imp} . Thus, K solves for the RRH_∞ C problem for P_{imp} , as claimed.

Before continuing with the formal proof, we give an intuitive argument which will clarify the role of the coupling conditions (1). Let us write P_{imp} as the stable system \bar{P} (stable because X is the stabilizing Riccati solution) with state-feedback $\xi = B_2' X x + u$, and close the controller loop with K as shown in Fig. 4. The controller K solves the robust regulation problem for P_{imp} , so sinusoids of frequency ω_k are rejected from any input of P_{imp} to v_2 . In particular, sinusoids are rejected from r to v_2 . Now we apply a unit amplitude sinusoid of frequency ω_k to input r . Since v_2 is asymptotically zero we have, in the frequency domain

$$0 = \bar{P}_{v_2, r} v + \bar{P}_{v_2, \xi} \xi.$$

In general, in order for this equation to be satisfied, ξ will have to have some nonzero amplitude, say γ_k . Since $v_1 = \xi$, therefore, v_1 will also have amplitude γ_k . Thus, $\|T_{v_1, r}\|_\infty \geq \gamma_k$. This argument goes through for every frequency that is regulated against, so the robust regulation constraint defines N lower bounds on $\|T_{v_1, r}\|_\infty$. The matrix inequalities (1) are satisfied only if all N lower bounds are less than γ . We first establish some notation before making these arguments precise. Open-loop transfer functions of \bar{P} will be denoted by the symbol \bar{P} with the appropriate subscripts. Thus, the open-loop transfer function from r to v_2 will be denoted by $\bar{P}_{v_2, r}$. Closed-loop transfer functions, i.e., transfer functions of the feedback system of Fig. 4, will be denoted in the usual way; thus $T_{v_1, r}$ denotes the closed-loop transfer function from r to v_1 .

Now we investigate the feedback system of Fig. 4 in more detail. In the frequency domain $\xi = T_{\xi, r}(s)r$, but since $v_1 = \xi$ (see Fig. 4), we may also write $\xi = T_{v_1, r}(s)r$. We also have $v_2 = \bar{P}_{v_2, r}(s)r + \bar{P}_{v_2, \xi}(s)\xi$, hence, substituting for ξ , $v_2 = \bar{P}_{v_2, r}(s)r + \bar{P}_{v_2, \xi}(s)T_{v_1, r}(s)r$. Thus

$$T_{v_2, r}(s) = \bar{P}_{v_2, r}(s) + \bar{P}_{v_2, \xi}(s)T_{v_1, r}(s).$$

Now let ω_k be a frequency that is regulated against. Because K solves the robust regulation problem for P_{imp} , we know that $T_{v_2, r}(j\omega_k) = 0$, hence

$$\bar{P}_{v_2, r}(j\omega_k) + \bar{P}_{v_2, \xi}(j\omega_k)T_{v_1, r}(j\omega_k) = 0. \quad (5)$$

We note that $\bar{P}_{v_2, r}(j\omega_k) = T_k B_1$ and $\bar{P}_{v_2, \xi}(j\omega_k) = T_k B_2$, where T_k is defined in ii) of the theorem statement. Thus (5) can be rewritten

as $T_k B_1 + T_k B_2 T_{v_1, r}(j\omega_k) = 0$, which implies

$$T_k B_2 T_{v_1, r}(j\omega_k) T_{v_1, r}(j\omega_k)^* B_2' T_k^* = T_k B_1 B_1' T_k^*. \quad (6)$$

From Theorem A.1 we know that P_{imp} cannot have transmission zeros from u to v_2 at the frequency $j\omega_k$. Thus, because transmission zeros are invariant under state-feedback, \bar{P} cannot have transmission zeros from ξ to v_2 , i.e., the matrix

$$\begin{bmatrix} j\omega_k I - \bar{A} & B_2 \\ [I \ 0] & 0 \end{bmatrix}$$

must have independent rows. This, in turn, implies that $T_k B_2$ has independent rows. In addition, because $\|T_{v_1, r}\|_\infty < \gamma$, we also know that $\gamma^2 I - T_{v_1, r}(j\omega_k) T_{v_1, r}(j\omega_k)^* > 0$. These two facts imply that $\gamma^2 T_k B_2 B_2' T_k^* > T_k B_2 T_{v_1, r}(j\omega_k) T_{v_1, r}(j\omega_k)^* B_2' T_k^*$. Now bring in (6) to conclude $\gamma^2 T_k B_2 B_2' T_k^* > T_k B_1 B_1' T_k^*$.

(ii \Rightarrow i). The first step is to show that

$$v^* (\gamma^{-2} L B_1 B_1' L' - L B_2 B_2' L') v < 0 \quad (7)$$

for any left eigenvector v of \bar{A} . With this in mind let v be a left eigenvector of \bar{A} and suppose that the corresponding eigenvalue is $j\omega_k$, so

$$v^* \bar{A} = j\omega_k v^*. \quad (8)$$

We recall from ii) of the theorem statement that L is defined by $L\bar{A} - \bar{A}L = B[I \ 0]$. Multiplying this equation through on the left by v^* , using (8), and then rearranging, we obtain $-v^* L = v^* B[I \ 0](j\omega_k I - \bar{A})^{-1}$. Using the definition of T_k this can be rewritten as

$$-v^* L = v^* B T_k. \quad (9)$$

Now from ii) of the theorem, we have $\gamma^{-2} T_k B_1 B_1' T_k^* - T_k B_2 B_2' T_k^* < 0$. Because $v^* B \neq 0$ (since, by Theorem A.2, the internal model (\bar{A}, B) is controllable), it follows that

$$\gamma^{-2} v^* B T_k B_1 B_1' T_k^* B' v - v^* B T_k B_2 B_2' T_k^* B' v < 0.$$

Hence, by (9), $\gamma^{-2} v^* L B_1 B_1' L' v - v^* L B_2 B_2' L' v < 0$, thus proving the claim.

Next, we bring a result of Scherer ([17, p. 128, Theorem 4]) to conclude that, by virtue of inequality (7), there exists a $W > 0$ such that

$$AW + W\bar{A}' + \gamma^{-2} L B_1 B_1' L' - L B_2 B_2' L' < 0.$$

Thus the controller K given in the theorem statement is well defined. Next, we show that this controller solves the \mathcal{H}_∞ problem for P . If the controller loop of G_{imp} is closed with K we obtain

$$T_{v_1, r} \sim \left[\begin{array}{c|c} \bar{A} - B_2 B_2' L' W^{-1} L & B_2 B_2' L' W^{-1} \\ \hline \bar{B}[I \ 0] & A \end{array} \middle| \begin{array}{c} B_1 \\ 0 \end{array} \right].$$

A change of coordinates via $\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$ results in

$$T_{v_1, r} \sim \left[\begin{array}{c|c} \bar{A} & -B_2 B_2' L' W^{-1} \\ \hline 0 & \bar{A} - L B_2 B_2' L' W^{-1} \end{array} \middle| \begin{array}{c} B_1 \\ L B_1 \end{array} \right]$$

which can be simplified by eliminating stable unobservable modes to yield

$$T_{v_1, r} \sim \left[\begin{array}{c|c} \bar{A} - L B_2 B_2' L' W^{-1} & L B_1 \\ \hline -B_2' L' W^{-1} & 0 \end{array} \right] =: \left[\begin{array}{c|c} \hat{A} & \hat{B} \\ \hline \hat{C} & 0 \end{array} \right].$$

We note that

$$\begin{aligned} & \Gamma' W^{-1} + W^{-1} A + \gamma^{-2} W^{-1} B B' W^{-1} + C' C \\ &= W^{-1} (A W + W A' + \gamma^{-2} L B_1 B_1' L' - I B_2 B_2' L') W^{-1} \\ &< 0 \end{aligned}$$

where the inequality is a consequence of (4). It follows from the bounded real lemma (the algebraic Riccati inequality analysis result—see for example [26, Lemma 2.2]) that K solves the \mathcal{H}_∞ problem for G_{imp} . Thus, again bringing in Redheffer ([9, Lemma 9]), K also solves the \mathcal{H}_∞ problem for P . In addition, since K has the internal model structure of Theorem A.2, K also solves the robust regulation problem for P ; hence the $\text{RRH}_\infty\text{C}$ problem for P .

C. Solution: Output Feedback Case

In this section the $\text{RRH}_\infty\text{C}$ problem (hence the RRNBU problem) is solved in the output feedback case. We make the following standard assumptions on P :

- OF1: (A, B_2) is stabilizable and (C, A) is detectable.
- OF2: (C_1, A, B_1) has no uncontrollable/unobservable modes on the imaginary axis.
- OF3: $D_1' C_1 = 0$ and $D_1' D_1 = I$.
- OF4: $B_1 D_1' = 0$ and $D_1 D_1' = I$.

Theorem 3.4 Let the plant P of Fig. 3 satisfy the standard assumptions OF1–OF4. Then the following are equivalent:

- i) There exists a controller admissible for P that solves the robust regulation problem from u to y at the frequencies $\omega_1, \dots, \omega_N$ while also making $\|T_{11}\|_\infty < \infty$ (the $\text{RRH}_\infty\text{C}$ problem).
- ii) There exists a controller admissible for P that renders $\|T_{11}\|_\infty < \infty$ and

$$I_k B' B' I_k' \succ (I_k \bar{B}_1 + [I_k 0])(I_k \bar{B}_1 + [I_k 0])' \quad (10)$$

for all $k = 1, \dots, N$. Here

$$\begin{aligned} I_k &= [I_k 0] \bar{C} (I_k \omega_k I - \bar{A})^{-1} \\ \bar{A} &= A + \gamma^{-2} C' C_1 + \gamma^{-2} C' C_1' Z - B' B' Z \\ \bar{B}_1 &= \gamma C' \\ Z &= \lambda (I - \gamma^{-2} \lambda \lambda)^{-1} \end{aligned}$$

λ is the positive semidefinite stabilizing solution to the algebraic Riccati equation

$$A' \lambda + \lambda A + \lambda (C_1 B_1' - B B') \lambda + C_1 C_1' = 0$$

and γ is the positive semidefinite stabilizing solution to the Riccati equation

$$A' \gamma + \gamma A' + \gamma (C_1 C_1' - C_2' C_2) \gamma + B_1 B_1' = 0$$

Moreover, if either (hence both) of these conditions hold, then a controller that solves the $\text{RRH}_\infty\text{C}$ problem for P is given by (11) found at the bottom of the page. Here I is the unique solution to $I \bar{A} - A I = B [I 0] \bar{C}$, $C_2 = C_2' (I + \gamma Z)$, the internal model matrices A and B satisfy 1)–4) of Definition 3.2, and W is a positive definite matrix satisfying the Lyapunov inequality

$$\begin{aligned} & W + W A' + \gamma^{-2} (I B_1 - B [I 0]) (I B_1 - B [I 0])' \\ & \quad - I B_2 B_2' I' < 0 \quad (12) \end{aligned}$$

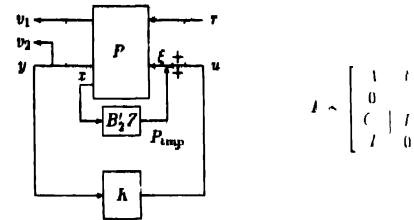


Fig. 5 The closed loop system consisting of plant P and controller K . P_{11} consists of stable system P with state feedback $B'Z$.

Proof ($i \Rightarrow ii$) This proof parallels that of the state feedback case. Suppose there exists a controller K that solves the $\text{RRH}_\infty\text{C}$ problem for P . The first step is to show that K also solves the $\text{RRH}_\infty\text{C}$ problem for

$$P_{11} \sim \left[\begin{array}{c|c} \bar{A} + B' B' Z & \bar{B}_1 - B \\ \hline B' Z & 0 \quad I \\ \hline C & I \quad 0 \end{array} \right]$$

where \bar{A} , \bar{B}_1 , and \bar{C} are defined in the theorem statement. To prove this, we make use of two system transformations as in [20].

Because there exists a controller K that solves the \mathcal{H}_∞ problem for P (and P satisfies assumptions OF1–OF4), the Riccati solutions λ and γ of the theorem statement exist. This follows from [9, Theorem 3]. In particular, γ exists. Therefore, by the dual of the Redheffer Lemma ([9, Lemma 9]), K also solves the \mathcal{H}_∞ problem for the plant

$$\left[\begin{array}{c|c} \bar{A} + \gamma C' C_1 & \bar{B}_1 - B \\ \hline C_1 & 0 \quad D_1 \\ \hline C & I \quad 0 \end{array} \right] \quad (13)$$

Similarly, because there exists a controller K that solves the \mathcal{H}_∞ problem for (13), we know that there exists a unique stabilizing solution $Z = 0$ to the λ Riccati equation for (13). In fact, it can be easily verified that $Z = \lambda (I - \gamma \lambda)^{-1}$. Thus, again bringing in Redheffer, K also solves the \mathcal{H}_∞ problem for

$$\left[\begin{array}{c|c} \bar{A} + \gamma C' C_1 + B_1 B_1' Z & \bar{B}_1 - B \\ \hline B' Z & 0 \quad I \\ \hline C_2 + \bar{B}_1' Z & I \quad 0 \end{array} \right]$$

which is precisely the plant P_{11} .

Next, we write P_{11} as the stable system P with state feedback $\xi = B' Z x + u$ and close the controller loop with K as shown in Fig. 5. The remainder of the proof then consists of an analysis of Fig. 5 just as the feedback system of Fig. 4 was analyzed in the state feedback case. The details are omitted.

($ii \Rightarrow i$) We show that K given by (11) solves the \mathcal{H}_∞ problem for P by showing that it solves the \mathcal{H}_∞ problem for P_{11} and then appealing to the Redheffer lemma ([9, Lemma 9]) and its dual Λ realization for the closed loop system consisting of P_{11} and K (stacking plant states on top of controller states) is shown in the

$$K \sim \left[\begin{array}{c|c} \bar{A} & 0 \\ \hline B_2 B_2' I' W^{-1} & \bar{A} + \gamma (C_1 C_1' - C_2' C_2) - B' B' Z - B_2 B_2' I' W^{-1} I \\ \hline B' L' W^{-1} & -B_2' Z - B' L' W^{-1} I \\ \hline & 0 \end{array} \right] \quad (11)$$

$$T_{11} \sim \left[\begin{array}{ccc|c} \bar{A} + B_2 B_2' Z & B_2 B_2' L' W^{-1} & -B_2 B_2' Z - B_2 B_2' L' W^{-1} L & \bar{B}_1 \\ \bar{B}[I_l 0] \bar{C}_2' & \bar{A} & 0 & \bar{B}[I_l 0] \\ \hline \bar{B}_1 \bar{C}_2 & B_2 B_2' L' W^{-1} & \bar{A} - \bar{B}_1 \bar{C}_2 - B_2 B_2' L' W^{-1} L & \bar{B}_1 \\ B_2' Z & B_2' L' W^{-1} & -B_2' Z - B_2' L' W^{-1} L & 0 \end{array} \right].$$

equation found at the top of the page. After changing coordinates via

$$\begin{bmatrix} I & 0 & 0 \\ L & -I & 0 \\ 1 & 0 & -I \end{bmatrix}$$

and eliminating stable unobservable and uncontrollable modes, we obtain

$$T_{11} \sim \left[\begin{array}{c|c} \bar{A} - L B_2 B_2' L' W^{-1} & L \bar{B}_1 - B[I_l 0] \\ \hline -B_2' L' W^{-1} & 0 \end{array} \right].$$

Now we can appeal to the bounded real lemma (as in the state-feedback case) to conclude that K solves the H_∞ problem for P_{imp} , and thus for P as well. But K also has the internal model structure of Theorem A.2, so K solves the $RRH_\infty C$ problem for P .

APPENDIX A THE ROBUST REGULATION PROBLEM

This appendix includes well-known results on the robust regulation problem that are used in the proofs of Section III. A precise definition of the robust regulation problem is given in the introduction, and all theorems here apply to the plant of Fig. 1. The first theorem addresses the question of when there exists a controller that solves this problem.

Theorem A.1 [7], [12], [11] There exists a controller that solves the robust regulation problem determined by $\omega_1, \dots, \omega_N$ for P if and only if

- i) (A, B_2) is stabilizable and (C_2, A) is detectable, and
- ii) the matrix $\begin{bmatrix} sI - A & B_2 \\ C_2 & 0 \end{bmatrix}$ has independent rows for every $k = 1, \dots, N$.

A controller K solves the robust regulation problem if and only if it contains the right "internal model." This "internal model principle" [1], [3], [5], [7], [8], [11], [12], [22], [23] is the central result in robust regulation. A particularly simple statement of this result is given in the following.

Theorem A.2 [1]: A controller K solves the robust regulation problem determined by $\omega_1, \dots, \omega_N$ for P if and only if

- i) K is admissible for P , and
- ii) K admits the realization

$$K \sim \left[\begin{array}{cc|c} A & 0 & \bar{B}[I_l 0] \\ * & * & A \\ \hline * & * & * \end{array} \right] \quad (14)$$

(the "*" entries are any matrices of appropriate dimensions) where

- a) $\text{spec}(A) = \{\pm j\omega_1, \dots, \pm j\omega_N\}$,
- b) every eigenvalue of \bar{A} has multiplicity l ,
- c) \bar{A} is diagonalizable, and
- d) (\bar{A}, \bar{B}) is controllable.

Matrices A and \bar{B} satisfying a)–d) of Theorem A.2 can be realized in many different ways. We give one realization here. Another may be found in, for example, [8].

With every frequency ω_k to be regulated against, associate system matrices \bar{A}_k and \bar{B}_k as follows. If $\omega_k = 0$, choose integrator dynamics

$$\bar{A}_k := 0 \in \mathbb{R}^{l \times l} \quad \text{and} \quad \bar{B}_k := I \in \mathbb{R}^{l \times l}.$$

If $\omega_k \neq 0$, choose harmonic oscillator dynamics

$$A_k := \begin{bmatrix} 0 & \omega_k I \\ -\omega_k I & 0 \end{bmatrix} \in \mathbb{R}^{2l \times 2l} \quad \text{and} \quad \bar{B}_k := \begin{bmatrix} 0 \\ I \end{bmatrix} \in \mathbb{R}^{2l \times l}.$$

Now set

$$A := \begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_N \end{bmatrix} \quad \text{and} \quad B := \begin{bmatrix} B_1 \\ \vdots \\ B_N \end{bmatrix}.$$

It is straightforward to verify that A and B as constructed do, in fact, satisfy requirements a)–d) of Theorem A.2.

The last result on robust regulation is a straightforward consequence of Theorem A.2.

Corollary A.3 Suppose K solves the robust regulation problem for P . Then K solves the robust regulation problem for any other finite-dimensional linear time-invariant plant for which it is an admissible controller.

REFERENCES

- [1] J. Abedor, "Robust regulation revisited," BCCIT Tech. Rep. 93-1, Dep. Mechanical Engineering, Univ. California, Berkeley, CA, 1993.
- [2] J. Abedor, K. Nagpal, P. Khargonekar, and P. Poolla, "Robust regulation with an H_∞ constraint," in *Control of Uncertain Systems*, Boca Raton, FL: CRC, 1991, pp. 95–110.
- [3] G. Bengtsson, "Output regulation and internal models—A frequency domain approach," *Automatica*, vol. 13, pp. 333–345, 1977.
- [4] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, June 1993 (draft).
- [5] L. Cheng and J. B. Pearson, "Frequency domain synthesis of multivariable linear regulators," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 3–15, 1978.
- [6] E. J. Davison, "The output control of linear time-invariant multivariable systems with unmeasurable arbitrary disturbances," *IEEE Trans. Automat. Contr.*, vol. AC-17, pp. 621–630, 1972.
- [7] E. J. Davison and A. Goldenberg, "Robust control of a general servomechanism problem: The servo compensator," *Automatica*, vol. 11, no. 5, pp. 461–471, 1975.
- [8] E. J. Davison, "The robust control of a servomechanism problem for linear time-invariant multivariable systems," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 25–34, 1976.
- [9] J. Doyle, K. Glover, P. Khargonekar, and B. Francis, "State-space solutions to standard H_2 and H_∞ control problems," *IEEE Trans. Automat. Contr.*, vol. AC-34, no. 8, pp. 831–847, 1989.
- [10] B. A. Francis, *A Course in H_∞ Control Theory (Lecture Notes in Control and Information Sciences, Vol. 88)*, New York: Springer-Verlag, 1987.
- [11] B. A. Francis and M. Vidyasagar, "Algebraic and topological aspects of the regulator problem for lumped linear systems," *Automatica*, vol. 19, pp. 87–90, 1983.
- [12] B. A. Francis and W. M. Wonham, "The internal model principle for linear multivariable regulators," *Appl. Math. Optimiz.*, vol. 2, no. 2, pp. 170–194, 1975.
- [13] S. Hara and T. Sugie, " H_∞ control problem with general boundary constraints," in *Proc. 28th IEEE Conf. Decision Contr.*, Tampa, FL, 1989.

- [14] P. P. Khargonekar, I. R. Petersen, and K. Zhou, "Robust stabilization of uncertain linear systems: Quadratic stabilizability and H_∞ control theory," *IEEE Trans Automat Contr*, vol. AC-35, pp. 356–361, 1990.
- [15] I. R. Petersen, "Disturbance attenuation and H_∞ optimization: A design method based on the algebraic Riccati equation," *IEEE Trans Automat Contr*, vol. AC-32, no. 5, pp. 427–429, May 1987.
- [16] C. Scherer, " H_∞ optimization without assumptions on finite and infinite zeros," *SIAM J Contr Optimiz*, vol. 30, pp. 143–166, 1992.
- [17] —, " H_∞ control by state feedback for plants with zeros on the imaginary axis," *SIAM J Contr Optimiz*, vol. 30, pp. 123–142, 1992.
- [18] T. Sugie and S. Hara, " H_∞ suboptimal control problem with boundary constraints," *Syst Contr Lett*, vol. 13, pp. 93–99, 1989.
- [19] T. Sugie and M. Vidyasagar, "Further results on the robust tracking problem in two degree of freedom control systems," *Syst Contr Lett*, vol. 13, pp. 101–108, 1989.
- [20] A. A. Stoorvogel, *The H_∞ Control Problem: A State Space Approach*. Hemel Hempstead: Prentice Hall International, 1992.
- [21] M. Vidyasagar, *Control Systems Synthesis*. Cambridge, MA: MIT Press, 1985.
- [22] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 3rd ed. New York, NY: Springer Verlag, 1985.
- [23] W. M. Wonham and J. B. Pearson, "Regulation and internal stabilization in linear multivariable systems," *SIAM J Contr Optimiz*, vol. 12, pp. 5–18, 1974.
- [24] Q. H. Wu and M. Mansour, "Robust output regulation for a class of linear multivariable systems," *Syst Contr Lett*, vol. 13, pp. 227–232, 1989.
- [25] G. Zames, "Feedback and optimal sensitivity: Model reference transformation, multiplicative seminorms and approximate inverses," *IEEE Trans Automat Contr*, vol. AC-26, pp. 301–320, 1981.
- [26] K. Zhou and P. P. Khargonekar, "An algebraic Riccati equation approach to H_∞ optimization," *Syst Contr Lett*, vol. 11, pp. 85–92, 1988.

Frequency-Domain Criteria of Robust Stability for Slowly Time-Varying Systems

A. Megretski

Abstract—The problem of stability of feedback systems with structured slowly time-varying uncertain gains is considered. For the case when the pair "uncertain gain/its derivative" belongs to a given convex set, a sufficient frequency domain condition of stability is obtained. This condition is an MIMO generalization of the SISO results derived in the 60s in the context of the positivity theory.

1. INTRODUCTION

Consider the system with uncertainty in the feedback loop shown in Fig. 1 where P is the linear time invariant (LTI) nominal plant and Δ is the linear block representing the uncertainty $\xi(t) = \delta(t)y(t)$. The (unknown) function $\delta(\cdot)$ is such that for any t the pair $(\delta(t), \dot{\delta}(t))$ belongs to a given set D of pairs of matrices. Note that in the case when $\delta_1 = 0$ for any $(\delta_0, \dot{\delta}_1) \in D$ the setup describes an LTI system with a real parametric uncertainty.

Manuscript received December 21, 1992; revised June 18, 1993 and April 1994. The original version of the manuscript was prepared when the author was visiting the Department of Electrical Engineering at the University of Newcastle, Australia. This work was supported in part by the Institute for Mathematics and Its Applications with funds provided by the National Science Foundation.

The author is with the Department of Electrical Engineering and Computer Engineering, Iowa State University, Ames, IA 50011 USA.
IEEE Log Number 9407030.

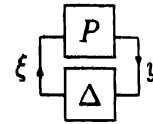


Fig. 1

The setup from Fig. 1 can be represented in the equivalent form

$$\dot{x}(t) = A(t)x(t) \quad A(t) = A_0 + B_1 \delta(t) C_1 \quad \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} I \\ C_0 \end{pmatrix} x(t) \quad (1.1)$$

where $x(t)$ is the state vector of P , $y = C_0 x$ and A, B, C are given matrices. A well studied case of the problem is when

$$A_0 = 0 \quad B_0 = C = I$$

$$D = \{(\delta_0, \dot{\delta}_1) \mid \|\delta_0\| \leq M, R(\dot{\delta}_1) \leq -\alpha \|\dot{\delta}_1\| - \epsilon\} \quad (1.2)$$

i.e. the pointwise bounds on the norms of $\dot{\delta}(t)$ and $\delta(t)$ and on the Lyapunov index of $\dot{\delta}(t)$ are given ($\|F\|$ and $R(F)$ denote the largest singular value and the largest real part of eigenvalue of F respectively, I denotes a unit matrix). It was shown in [1] that system (1.1)–(1.2) is stable provided that $\epsilon < \epsilon_0(M, \alpha)$ is sufficiently small. Explicit bounds $\epsilon_0(M, \alpha)$ were obtained for example in [2] and [3] where a more general formulation is also considered with M, α and ϵ being the average values of $\|\dot{\delta}(t)\|$, $R(\dot{\delta}(t))$ and $\|\delta(t)\|$ respectively. See also [4] for the infinite dimensional case and see [5] for the application of the technique in the gain scheduling and adaptive control. The general results mentioned above establish the so called frozen time principle in the analysis of slowly time varying (STV) systems. The issues of concern, however, are the conservativity of the upper bound $\epsilon < \epsilon_0(M, \alpha)$ and the complexity of the analysis required to check the condition $R(\dot{\delta}(t)) \leq -\alpha$. First, most practical systems of interest do have structure which means that $\dot{\delta}(\cdot)$ is not expected to be an arbitrary matrix function satisfying the bounding conditions (1.2). For example, when the order of the plant P on Fig. 1 is greater than one, the coefficients of $\dot{\delta}$ are not independent. Since the stability condition $\epsilon < \epsilon_0(M, \alpha)$ does not utilize such information, the bound may be very conservative. Although no systematic study of the conservativity is delivered in the note, a reasonable example of a second order system is given for which the criterion from Theorem 1 below (which does utilize the structure) gives about 30 000 times larger bound on the admissible rate of variation. Second, when a generic set of expected values of matrices $\dot{\delta}(t)$ is given, the problem of finding the upper bound α in (1.2) is equivalent to a robust stability problem in the presence of a multivariable real parametric uncertainty which is known to be NP hard.

An alternative approach to studying STV systems was developed within the positivity theory (see for example [6] and [7] and references therein). In many (but not in all) examples, this approach yields much (thousands times) less conservative upper bounds than the unstructured one. However, the classical papers like [6] and [7] treat the case of a single time varying parameter only. In this note, relatively simple sufficient frequency domain conditions of robust stability in the STV case are proved. With some reservations, the result can be viewed as a generalization of the criteria [6]–[7] to the case of any number of time varying coefficients. The analysis required by the new criteria is similar to that used in the real μ case (see for example [8]) where the uncomputable real μ is replaced by a convenient upper bound. A similar approach to the analysis of STV systems was developed independently in [9].

II. MAIN RESULT

We consider system equations given by (1.1) where A_0 is a Hurwitz $n \times n$ matrix and D is a given convex compact set of pairs of $m \times l$ matrices containing the pair of zero matrices. For a complex matrix F , let F^* denote the Hermitian conjugate. We shall write $F \geq 0$ if the real part of $f^* F f$ is not negative for any complex vector f .

Definition 1: The uncertain STV system (1.1) is called stable (exponentially) if there exist $c_1, c_2 > 0$ such that $|x(t_1)| \leq c_1 e^{c_2(t_1-t_0)} |x(t_0)|$ for any $x(\cdot)$, $\delta(\cdot)$ satisfying (1.1) and for any $t_1 > t_0$.

Definition 2: System (1.1) is called Ψ -stable if

- $\delta_1 = 0$ for any $(\delta_0, \delta_1) \in D$ [i.e., $A = \text{const}$ in (1.1)];
- there exist $c > 0$ and an $l \times l$ transfer matrix Ψ (bounded in the right-half plane) such that

$$[I - \delta_0 G(j\omega)] \Psi(j\omega) - cI \geq 0 \quad \forall (\delta_0, \delta_1) \in D, \omega \in \mathbb{R}$$

where $G(s) = C_0(sI - A_0)^{-1}B_0$.

It is well known and easy to show (and follows from the proof of Theorem 1 below) that Ψ -stability implies the ordinary exponential stability. The advantage of using the Ψ -stability is in the relative simplicity of its verification. A possible disadvantage is that a stability criterion based on the Ψ -stability may be conservative. However, in many applications, the Ψ -stability is "close" to the standard robust stability. For example, if $m = 1$, i.e., B_0 is a vector, then Ψ -stability is equivalent to robust stability (see [10] and [11]).

Let us formulate the main result of this note. For $r \in (0, -\mathbf{R}(A_0))$, the modified system (1.1) is defined by

$$A_1 = A_0 + rI, \quad B_1 = B_0, \quad C_1 = C_0,$$

$$D_1 = \left\{ \left(\delta_0 - \frac{\delta_1}{2r}, 0 \right) : (\delta_0, \delta_1) \in D \right\}$$

where A_1, B_1, C_1, D_1 are the modified A_0, B_0, C_0, D .

Theorem 1: System (1.1) is stable if the modified system (1.1) is Ψ -stable, i.e., if there exists $c > 0, r \in (0, -\mathbf{R}(A_0))$, and an $m \times m$ rational function Ψ (bounded in the right-half plane $\text{Re}(s) > 0$) such that

$$\left[I - \left(\delta_0 - \frac{\delta_1}{2r} \right) G(j\omega - r) \right] \Psi(j\omega - r) - cI \geq 0 \quad (2.1)$$

for any $\omega \in \mathbb{R}, (\delta_0, \delta_1) \in D$.

Remark 1: For a fixed $r > 0$, condition (2.1) is convex in $\Psi(\cdot)$. The following procedure can be used to apply Theorem 1 in robustness analysis.

a) choose a finite set $\{c_0, c_1, \dots, c_q\}$ of scalar rational functions (bounded in the right-half plane);

b) for a fixed $r > 0$, use standard convex optimization to check whether there exists $\Psi(s) = \sum_{j=0}^q \Psi_j \Phi_j(s)$ (where Ψ_j are constant real matrices) satisfying (2.1) (for some $c > 0$, for any $\omega \in \mathbb{R}, (\delta_0, \delta_1) \in D$);

c) use random search over r to find whether the function Ψ in b) exists for some $r \in (0, -\mathbf{R}(A_0))$. It can be shown that, when D is a polytope and r is fixed, the search for Ψ_j can be reduced to solving a system of Linear Matrix Inequalities (LMI's), see [12].

Remark 2: Consider the problem of robust stabilization of the uncertain STV system on Fig. 1. This is the case when the transfer matrix G of P is not fixed but has to be chosen from an affine set given by the Youla parameterization $G = G_0 + G_1 H G_2$, where G_0, G_1, G_2 are given stable transfer matrices, and H ranges over the set of stable transfer matrices of appropriate dimension. When $G_2 = I$ (i.e., in the state feedback stabilization problem), the set of H , for which the robust stability of the STV system can be proved by using Theorem 1 with a fixed r , allows a convex parameterization, which reduces the problem to solving a

system of LMI's. This parameterization is similar to the projective parameterization introduced in [10]:

$$H(s) = \Phi(s) \Psi(s+r)^{-1}$$

where (Ψ, Φ) is the "parameter"—arbitrary pair of stable transfer matrices of appropriate size such that

$$\Psi(j\omega) - \left(\delta_0 - \frac{\delta_1}{2r} \right) [G_0(j\omega - r) \Psi(j\omega) + G_1(j\omega - r) \Phi(j\omega - r)] \geq cI$$

for some $c > 0$ for any $\omega \in \mathbb{R}, (\delta_0, \delta_1) \in D$.

III. THE "RANK ONE" CASE

In this section, we discuss application of the main result in the case when $m = 1$, i.e., ξ is scalar. Note that this is not the case of one single uncertain coefficient: basically, this is the case when we investigate stability of the linear differential equation

$$u^{(k)} + f_{k-1}(t)u^{(k-1)} + \dots + f_0(t)u = 0 \quad (3.1)$$

where u is scalar. The following result is a simple corollary to Theorem 1, based on the fact [10], [11] that the Ψ -stability is equivalent to the ordinary exponential stability in the rank one case ($m = 1$).

Theorem 2: Let K be a convex compact subset of \mathbb{R}^{2k} such that $(\bar{f}_0, \bar{f}_1, \dots, \bar{f}_{k-1}, 0, 0, \dots, 0) \in K$ for some Hurwitz polynomial $f(s) = f_0 + f_1 s + \dots + f_{k-1} s^{k-1} + s^k$. The differential equation (3.1) with the time-varying coefficients f_i such that

$$(f_0(t), f_1(t), \dots, f_{k-1}(t), \dot{f}_0(t), \dot{f}_1(t), \dots, \dot{f}_{k-1}(t)) \in K$$

for any t , is stable if for some $r > 0$ any polynomial

$$p_r(s) = (s-r)^k + \left(\bar{f}_{k-1} - \frac{\bar{g}_{k-1}}{2r} \right) (s-r)^{k-1} + \dots + \left(f_0 - \frac{\bar{g}_0}{2r} \right) \quad (3.2)$$

with $(\bar{f}_0, \bar{f}_1, \dots, \bar{f}_{k-1}, g_0, g_1, \dots, g_{k-1}) \in K$, is a Hurwitz polynomial.

Consider, as an example, the second-order differential equation

$$\ddot{w}(t) + f_1(t)\dot{w}(t) + f_2(t)w(t) = 0 \quad (3.3)$$

where $w(t)$ is scalar, and $f_j(t)$ are uncertain coefficients. Assume that $f_j(t)$ are within 100% of their (known) minimal values $f_1 = 1$ (s^{-1}) and $f_2 = F \geq 2$, i.e.,

$$1 \leq f_1(t) \leq 2, \quad F \leq f_2(t) \leq 2F. \quad (3.4)$$

Our aim is to find admissible relative rate of variation a for $f_j(t)$, i.e., to find $a > 0$ (as large as possible) such that the conditions

$$|\dot{f}_1(t)| \leq a, \quad |\dot{f}_2(t)| \leq Fa \quad (3.5)$$

and (3.4) guarantee exponential stability of (3.3).

First, let us apply Theorem 2 of this note in the case of the system setup (3.3)–(3.5). The sufficient condition of stability is existence of $r > 0$ such that all the polynomials

$$p(s) = (s-r)^2 + \left(\bar{f}_1 - \frac{\bar{g}_1}{2r} \right) (s-r) + \left(\bar{f}_2 - \frac{\bar{g}_2}{2r} \right),$$

$$\bar{f}_1 \in [1, 2], \quad |\bar{g}_1| \leq a, \quad \bar{f}_2 \in [F, 2F], \quad |\bar{g}_2| \leq Fa$$

are Hurwitz. It is easy to see (remember that $F \geq 2$) that the condition is satisfied for $r = 1/4, a < 1/4$. Therefore, $a < 1/4$ is a criterion of stability given by Theorem 2.

Now let us apply the result [2] (Theorem 3.2, (ii), with corrections),

$$\epsilon(M, \alpha) = \frac{1}{2} \frac{1}{(N-1)^2} \left(1 - \frac{1}{2N-1} \right)^{1/N-2} \frac{\alpha^{1/N-2}}{(2MF)^{1/N-1}} \quad (3.6)$$

in analysis of the system (3.3)–(3.5). To do this, we should represent (3.3) in a state-space form (1.1). One can use the canonical form, or any equivalent state-space realization,

$$\dot{A}(t) = \begin{bmatrix} 0 & 1 \\ -f_2(t) & -f_1(t) \end{bmatrix} \quad \text{or}$$

$$\dot{A}(t) = S \begin{bmatrix} 0 & 1 \\ -f_2(t) & -f_1(t) \end{bmatrix} S^{-1}$$

Indeed, we have

$$\operatorname{Re}(\lambda(t)) \geq -0.5 \operatorname{tr}(\dot{A}(t)) = -0.5$$

$$\|\dot{A}(t)\| \geq \sqrt{|\det \dot{A}(t)|} = \sqrt{2F} \quad \|\dot{A}(t)\| \geq \sup |f_1(t)| = a$$

Hence, the criterion would not give a bound better than $a < 1/2\sqrt{2}F$. Even for the smallest value of F under consideration $F = 2$, this is about 30000 times worse than the criterion $a < 1/4$. Also, the last bound yields $a \rightarrow 0$ as $F \rightarrow \infty$ while the bound given by Theorem 2 does not depend on $F \in [2, \infty)$. Therefore, for the example under consideration the criterion [2] does not give good practical numbers, and also does not allow us to observe the important theoretical phenomenon: the admissible bound for the relative rate of variation does not tend to zero as the phase of the nominal system poles tends to $\pm\pi/2$.

IV. CONCLUDING REMARKS

The sufficient criterion of stability presented in this paper is far from being necessary. It appears that the criterion would provide a good balance between complexity and conservativity of analysis for dynamical systems with 'low to medium' speed time-varying coefficients (i.e., not too fast). It can be shown that the criterion is based on using a certain set of Integral Quadratic Constraints (see [13]) describing STV gains. Therefore, it can be used in the case of a mixed uncertainty combining real parametric uncertainty, fast time-varying gains, conic nonlinearities, unmodeled IT dynamics, and so on. These extensions will be discussed in a forthcoming paper.

V. APPENDIX

Proof of Theorem 1. In this proof, L_2 denotes the standard Hilbert space of square summable functions $h: \mathbf{R} \rightarrow \mathbf{R}^l$ with the norm $\|\cdot\|$ and the scalar product (\cdot, \cdot) . If δ is a bounded measurable matrix-valued function on \mathbf{R} , M_δ denotes the operator of multiplication by δ , i.e., $(M_\delta h)(t) = \delta(t)h(t)$, acting from one L_2 -space to (generally) another. Similarly, if $F = F(s)$ is a rational matrix-valued function bounded on the imaginary axis, the L_F denotes the Fourier operator of multiplication in the frequency domain: $h_1(j\omega) = F(j\omega)h(j\omega) \forall \omega \in \mathbf{R}$ for $h_1 = T_F h$ where T_F denotes the Fourier transform. Remember that $T_F L_2 = L_2$ and that if F is bounded in the right-half plane $\operatorname{Re}(s) > 0$ then the operator T_F is causal, i.e., $(T_F h)(t) = 0$ if $h(t_1) = 0$ for any $t_1 < t$.

It is sufficient to show that the operator $I - M_\delta L_2$ is invertible for any $\delta(\cdot)$, and that the norm of $(I - M_\delta L_2)^{-1}$ is bounded. Since the set of $\delta(\cdot)$ is convex and contains the zero function, one only needs to show that there exists $q > 0$ such that

$$\|\xi - M_\delta T_F \xi\| \geq q \|\xi\| \quad (4.1)$$

for any δ and for any $\xi \in L_2$. Note that since $(0, 0) \in \mathcal{D}$, condition (2.1) implies that $\Psi(j\omega) - qI \geq 0$ for any $\omega \in \mathbf{R}$. Therefore, since Ψ is bounded in the right-half plane, $\Psi(s) - qI \geq 0$ for any s with $\operatorname{Re}(s) \geq 0$. Hence, the rational function $\Psi(s)^{-1}$ where $\Psi(s) = \Psi(s+j)$ is well defined and bounded in the right-half plane. Therefore, the operator T_Φ is invertible and it is sufficient to prove

existence of $q_1 > 0$ such that

$$\|I_\Phi \xi - M_\delta L_2 \Phi \xi\| \geq q_1 \|\xi\| \quad (4.2)$$

for all $\delta(\cdot)$ and for all $\xi \in L_2$. We will prove this by showing that $\|(I - T_\Phi \xi - M_\delta L_2 \Phi \xi)\| \geq \epsilon \|\xi\|^2$.

Suppose (with no loss of generality) that ξ has compact support. Let $u = I_\Phi \xi$, $y = T_\Phi \xi$, $v(t) = \epsilon^{-1} \xi(t) - u(t)$, $y(t) = \epsilon^{-1} y(t)$. It is easy to check that u , y are square summable, and $u = \Psi v$, $y = \Gamma v$ where $\Gamma(s) = \Gamma(s-j)$. Therefore, by (2.1), using the Parseval formula, we have $(v, u) = (\delta_0 - \delta_1/(2j))y_1 \geq \epsilon \|\xi\|^2$ for any $(\delta_0, \delta_1) \in \mathcal{D}$. Since the operators T_Φ and $T_\Phi \Phi$ are causal and the last inequality holds for an arbitrary $\xi \in L_2$, we have $\operatorname{tr}\{(2jD(t) - D(t))W(t)\} > 0$ for any t where

$$D(t) = [I - \delta(t) - \epsilon I] \quad W(t) = \begin{bmatrix} u(t) \\ y(t) \\ v(t) \end{bmatrix}$$

$$W(t) = \int_{-\infty}^t (\tau) v(\tau)^* \epsilon^{-1} \tau d\tau$$

Note that the matrix W is bounded and $W(\tau) = 0$ for τ approaching $-\infty$. Hence

$$\begin{aligned} & \|(I - T_\Phi \xi - M_\delta L_2 \Phi \xi) - \epsilon \xi\| \\ &= \operatorname{tr} \int D^{-1} u^* dt - u \int \epsilon^{-1} D W dt \\ &= \int \epsilon^{-1} u^* \{(2jD - D)W\} dt \geq 0 \end{aligned}$$

(where the integrals are over \mathbf{R}) \square

ACKNOWLEDGMENT

The author is grateful to the anonymous referees for their comments.

REFERENCES

- [1] C. A. Desoer, 'Slowly varying system $\dot{x} = A(t)x$ ', *IEEE Trans Automat Contr*, vol. AC-14, pp. 780–781, 1969.
- [2] A. Ilchmann, D. H. Owens, and D. Pratzel Wolters, 'Sufficient conditions for stability of linear time-varying systems', *Syst Contr Lett*, vol. 9, pp. 157–163, 1987.
- [3] V. Solo, 'On the stability of slowly time-varying linear systems', preprint, 1992 (submitted to *MCSS*).
- [4] M. Dahleh and M. A. Dahleh, 'On slowly time-varying systems', *Automatica*, vol. 27, no. 1, pp. 201–205, 1991.
- [5] E. W. Kamien, P. P. Khargonekar, and A. Lannanbaum, 'Control of slowly time-varying linear systems', *IEEE Trans Automat Contr*, vol. AC-34, pp. 1283–1285, 1989.
- [6] M. Friedman and G. Zames, 'Logarithmic variation criteria for the stability of systems with time-varying gains', *SIAM J Contr*, vol. 6, no. 3, pp. 487–507, 1968.
- [7] M. K. Sundareshan and M. A. L. Thathachar, 'L₂-stability of linear time-varying systems: Conditions involving noncausal multipliers', *IEEE Trans Automat Contr*, vol. AC-17, pp. 504–510, 1972.
- [8] M. K. H. Fan, A. I. Tits, and J. C. Doyle, 'Robustness in presence of mixed parametric uncertainty and unmodeled dynamics', *IEEE Trans Automat Contr*, vol. 36, pp. 25–38, 1991.
- [9] S. Dasgupta, G. Chockalingam, B. D. O. Anderson, and M. Fu, 'Lyapunov functions for uncertain systems with applications to the stability of time-varying systems', submitted to *IEEE Trans Circuits Syst*.
- [10] A. Rantzer and A. Megretski, 'A convex parameterization of robustly stabilizing controllers', Tech. Rep. IRISA/MAT 92-0024, Roy. Inst. Technol., 1992, 12 p.
- [11] B. D. O. Anderson, S. Dasgupta, P. Khargonekar, I. J. Kraus, and M. Mansour, 'Robust strict positive realness: Characterization and construction', *IEEE Trans Circuits Syst*, vol. 37, pp. 869–876, 1990.
- [12] S. Boyd and I. F. Ghaoui, 'Method of centers for minimizing generalized eigenvalues', preprint available via anonymous ftp to: stanford.edu in pub/boyd/r/ports/gevc.ps.7.
- [13] A. Megretski, 'Power distribution approach in robust control', in *Proc 12th IFAC World Congr.*, Sydney, Australia, 1993, pp. 399–402.

Routing and Scheduling in Heterogeneous Systems: A Sample Path Approach

Panayotis D. Sparagkis

Abstract—Consider the problem of routing customers to a set of K parallel servers that have different rates. Each server has a buffer with infinite capacity. The arrival process is general and the service times are assumed to be i.i.d. exponential random variables. Using sample path arguments, we show that, given any Bernoulli policy π , there exists another policy ρ which outperforms π by partially using a randomized version of a round-robin policy. Moreover, ρ is easily specified and implemented. We present extensions of our results to systems with finite capacities and service times that have an increasing hazard rate. Finally, a similar result is shown to hold in the context of scheduling customers from a set of K parallel queues.

1. INTRODUCTION

Consider a queueing system with K parallel queues, each of which has a buffer with infinite or finite capacity and a dedicated server. There is a single arrival stream and a controller that routes customers to one of the K queues as soon as they arrive. The situation arises in at least two contexts: communication networks and multiprocessor computer systems. In the context of computer networks, for example, there are usually multiple paths to remote destinations and routing decisions have to be made locally, i.e., at the source node, based on information regarding the delay characteristics of the available paths.

The determination of a routing policy that is optimal with respect to a certain performance criterion (such as throughput or delay) has been the subject of extensive research in the past. In the simple case, i.e., when service times are i.i.d. exponential random variables and the controller has queue length information available at arrival instants, the shortest queue (SQ) policy has been shown to maximize throughput in systems with infinite capacities (see [3], [19], and [21]). This is often referred to as the symmetric routing problem. In addition, it has been shown that the SQ policy stochastically minimizes the number of customers that are lost by any time t when the buffers have finite queueing space, e.g., [5], [9], [16]. A counterexample for nonexponential service times is provided in [20]. Very recently, the optimality of the SQ policy was shown to hold over all service time distributions that are increasing in likelihood ratio [17]. When the rates at the K servers are not equal, however, the problem becomes significantly more difficult (see [1], [2]). A routing policy which for an infinite number of stations performs almost optimally over the class of deterministic policies is described in [11].

In an interesting variation of the symmetric routing problem, the controller does not have any queue length information. Indeed, this is a more plausible assumption in many applications. It has been shown that, when the servers are homogeneous and the customers have either exponential [3] or deterministic [12], [13] service times, the round-robin (RR) policy that alternates among the queues maximizes the system's throughput. Recently, it was proven in [6] that, in fact, the optimality of the RR policy covers all service-time distributions that have an increasing hazard rate. Extending the basic symmetric routing problem, it was assumed in [13] that there exists multiple arrival streams, each of which has a controller that selects one of the K queues each time a customer has to be routed, based only on its own

previous routing decisions. When the service times are deterministic and all the servers operate at the same rate, it was shown in [13] that a set of randomized RR policies (see Section II for a definition) becomes an equilibrium set of policies. That is, if all controllers use a randomized version of RR then, from the perspective of minimizing the expected steady-state delay of customers on its own stream, no individual controller has incentive to deviate from this scheme.

When the servers operate at different rates, a simple routing policy is to use Bernoulli splitting, i.e., route a customer to queue i with probability p_i , $\sum_{i=1}^K p_i = 1$, regardless of past decisions. The idea is to allocate larger portions of the incoming traffic to faster servers, effectively aiming to materialize the most benefits of load distributing. It is the main result of this note that if the service times are exponential then, given any Bernoulli policy π , there exists another routing policy, which partly uses a randomized version of RR, and which is superior to π in the sense of separable increasing convex ordering (the definition given in the following) of the queue length process. This will be proved in Section IV. Note that π need not be the optimal Bernoulli policy. As it will be seen, given the routing probabilities for π , policy ρ can be easily specified and implemented. Thus, this result has immediate applications since, in practice, probabilistic splitting is a common policy for static routing. When the system is Markovian, a numerical example shows that ρ can significantly reduce the stationary mean system delay at moderate-to-high traffic.

A similar result is shown to hold in the context of scheduling customers from K queues, each of which receives customers from a dedicated arrival stream (see Section V). Here, it is assumed that there is a single server which may switch to one of the queues according to a scheduling policy that is determined by a controller. The server can be thought as a communication channel shared by a number of stations. In this case, our result covers probabilistic collision-free access policies that are implemented in a distributed fashion. Another application arises in the context of scheduling transmissions in packet radio networks, where time slots are assigned to various classes of customers satisfying certain interference constraints (see, for example, [4] and [15]). When scheduling is done probabilistically, the policy is often referred to as random polling.

II. PRELIMINARIES

Let \mathbf{q}, \mathbf{r} , be two vectors in \mathbb{R}^K . We introduce the notation q_k to denote the k th largest element in vector \mathbf{q} and define the following ordering (see [8]).

Definition 1: Vector \mathbf{q} is said to weakly majorize vector \mathbf{r} (written $\mathbf{r} \prec_w \mathbf{q}$) if

$$\sum_{i=1}^k r_i \leq \sum_{i=1}^k q_i, \quad k = 1, \dots, K.$$

In particular, if $\sum_{i=1}^K r_i = \sum_{i=1}^K q_i$, then \mathbf{q} is said to majorize \mathbf{r} . The ordering in Definition 1 is usually referred to as weak submajorization (as opposed to the weak supermajorization ordering; see [8] for a comprehensive treatment). Throughout the rest of the note, we will use the broad term majorization to refer to the ordering " \prec_w ." Majorization has been a useful analysis tool in the context of routing or scheduling of customers (see, for example, [7] and [9]). The ordering makes precise the idea that the vectors of queue lengths in one system are "less" or "more balanced" than those in another system, thus, setting a convenient framework for making sample path comparisons among different load distributing policies.

Following the notation in [8] let C_1^1 denote the class of functions $f: \mathbb{R}^K \rightarrow \mathbb{R}$ of the form $f(\mathbf{q}) = \sum_{i=1}^K g(q_i)$, where $g: \mathbb{R} \rightarrow \mathbb{R}$

Manuscript received March 19, 1993; revised February 20, 1994. This work was supported in part by an IBM Graduate Fellowship Award.

The author was with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA and is now with CS First Boston Investment Management Group, New York, NY 10022 USA.

IEEE Log Number 9407032.

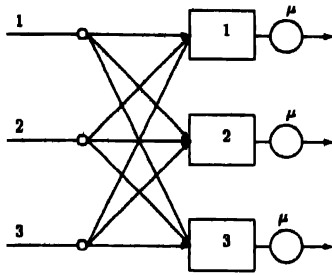


Fig. 1 An example of a system with multiple arrival streams

is increasing and convex C_1 is a subset of the class of Schur convex functions defined on \mathbb{R}^K (see [8]) thus having the property $r \prec q \Rightarrow f(r) \leq f(q) \forall f \in C_1$. A natural way to associate C_1 with a stochastic majorization ordering between random vectors is the following (see [8] for a complete discussion of stochastic majorization)

Definition 2 If N and M are random vectors we have

$$M \leq_{C_1} N \text{ if } F[\phi(M)] \leq F[\phi(N)] \quad \forall \phi \in C_1$$

The aforementioned ordering is sometimes referred to as separable increasing convex ordering ([6]). Finally we write $\{N(t) : t \geq 0\} \leq_{C_1} \{M(t) : t \geq 0\}$ if

$$(N(t_1), \dots, N(t_n)) \leq_{C_1} (M(t_1), \dots, M(t_n))$$

for all n, t_1, \dots, t_n

III THE RANDOMIZED ROUND ROBIN POLICY

In this section we consider a system of K parallel queues that admit customers from M external arrival streams. Each queue has infinite capacity and a dedicated server. Customers from each stream arrive at a controller that routes them to one of the K queues (see Fig. 1). The controller has available the past routing decisions on its own stream but no queue length information. We assume that the service times at each queue form a sequence of i.i.d. exponential random variables with a constant rate μ . On the other hand the sequence of arrival times on any particular stream is assumed to be independent of all service times but otherwise arbitrary. For example it can be a deterministic sequence. The arrival processes are mutually independent although not necessarily identically distributed.

We are interested in the class of symmetric routing policies Σ defined as follows. A policy belongs to Σ if for each arrival stream the probability that the i th customer is routed to queue j is $1/K$ for all $j = 1, \dots, K, i = 1, 2, \dots$. An example is a policy that uses Bernoulli splitting with equal routing probabilities at each controller. Another example is a policy that uses a randomized round robin (R3) policy (introduced in [13]) at each controller. Recall that given the first K customers to arrive a round robin (RR) policy routes exactly one customer to each of the K queues according to some routing pattern and repeats this pattern thereafter. R3 can be thought of as a stochastic version of RR: the controller uses a round robin policy that selects the routing pattern for the first K customers randomly in such a way that the probability of the i th customer being routed to queue j is $1/K, i, j = 1, \dots, K$. An example is a policy that selects a routing pattern among the K cyclic shifts of the set $\{1, \dots, K\}$ with probability $1/K$. Given a policy π in Σ let $N_i^\pi(t)$ denote the number of customers in queue i at time t and $N^\pi(t) = (N_1^\pi(t), \dots, N_K^\pi(t))$. Finally let $D^\pi(t)$ denote the total number of departures by time t under π .

Using dynamic programming it was shown in [13] that in a system that consists of K queues with deterministic servers and

M independent but identically distributed arrival streams, a set of R3 policies employed by the M controllers is optimal among all policies in the following sense: if each of $M-1$ controllers uses a R3 policy then the M th controller should do the same. The result shows that for the system considered in this section a set of R3 policies minimizes the queue length process in the sense of separable increasing convex ordering, among all policies in Σ . The proof follows the steps taken in [13].

More specifically as it will become clear in the following we will prove the optimality of a set of R3 policies: it suffices to show that RR is optimal among all policies that use no queue length information when $M = 1$ and in addition there exist K independent arrival streams one to each queue, that are identically distributed. Let Σ_1 be the class of feasible routing policies for this system. The result can be shown by constructing a modified probability space in which all arrivals from the K dedicated streams occur simultaneously at the K queues and all departures from the K queues also occur simultaneously at the jump points of a Poisson process with parameter μ . Given a policy $\pi \in \Sigma_1$ let $\{N^\pi(t) : t \geq 0\}$ denote the queue length process in this modified space. Clearly $N^\pi(t) = N^\pi(t)$ for all $t \geq 0$ and $i = 1, \dots, K$ but $\{N^\pi(t) : t \geq 0\} \neq \{N^\pi(t) : t \geq 0\}$. That is the joint statistics of the queue lengths are modified.

A slight extension of an argument given in [3] (see also [18] for more details) shows that $N^{R3}(t) \leq N^\pi(t)$ for all $\tau \in \Sigma_1, t \geq 0$. Then by an elementary property of majorization

$$(N^{R3}(t_1), \dots, N^{R3}(t_n)) \leq_{C_1} (N^\pi(t_1), \dots, N^\pi(t_n)) \quad \forall n, t_1, \dots, t_n$$

Since the marginal distributions of the queue lengths are not modified by the construction it follows that $\sum_{i=1}^K \sum_{j=1}^M F[q(N_i^\pi(t))]$ is increasing and convex. Thus we get the following proposition.

Proposition 1 In a system with K dedicated arrival streams and one common external stream RR minimizes the queue lengths in the following sense

$$\{N^{R3}(t) : t \geq 0\} \leq_{C_1} \{N^\pi(t) : t \geq 0\}$$

for all π in Σ_1 provided that $N^{R3}(0) = N_j^\pi(0)$ for all $j = 1, \dots, K$.

Going back to our original problem it now becomes easy to see that a set of R3 policies employed by the M controllers outperforms all policies in Σ . Let ρ be a policy in Σ such that on at least one stream say stream 1 a policy different from R3 is used. The key is in noting that due to the way Σ is defined the aggregate arrival process from streams 2 to M is such that each queue sees an identically distributed subsequence of arrivals. In other words conditioned on arrivals that are not from the first stream the arrival process at queue i is equal in law to that at queue j for all i and j . Therefore it follows from Proposition 1 that a policy ρ which copies ρ on all streams except from the stream 1 where it applies R3 instead does outperform ρ in the sense of separable increasing convex ordering of the queue length process. Repeating the argument for all streams on which ρ does not use R3 we get the following result.

Proposition 2 In a system with M common external arrival streams a set of R3 policies minimizes the queue lengths in the following sense

$$\{N^{R3}(t) : t \geq 0\} \leq_{C_1} \{N^\pi(t) : t \geq 0\}$$

for all π in Σ provided that $N^{R3}(0) = N_j^\pi(0)$ for all $j = 1, \dots, K$.

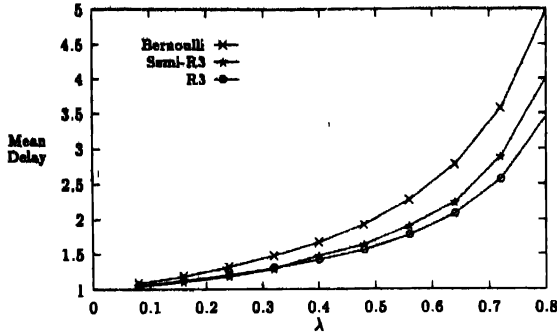


Fig. 2. A comparison of three policies.

Remark 1: As a consequence of the aforementioned result, it follows that a set of R3 policies maximizes the expected value of $D^*(t)$, for all $t \geq 0$, among all policies in Σ . Note, however, that this set of policies need not be optimal over all feasible routing policies (in particular, nonsymmetric policies). For example, $K = M$ and the arrival processes are identically distributed renewal processes with deterministic interarrival times, then the optimal policy is the one that assigns a single queue to each stream, so that all customers from that stream are routed to one particular queue. It follows that this scheme reduces to an RR policy applied to the aggregate arrival process (streams 1 to M), which by Proposition 1 ensures optimality.

A numerical example is given in Fig. 2. There are four Poisson streams, each with rate λ , and four queues, each with an exponential server of rate $\mu = 1$. The graph compares the steady-state mean system delay of a set of Bernoulli policies against that of a set of R3 policies as well as a set of semi-R3 policies. The latter is defined to be one in which each of streams 1 and 2 apply an R3 algorithm to queues 1 and 2, and each of streams 3 and 4 apply an R3 algorithm to queues 3 and 4. We assume that all Bernoulli policies use a vector of equal routing probabilities, which then makes the system equivalent to a set of four independent $M/M/1$ queues. The performance of the R3 and semi-R3 policies was determined via a simulation experiment. Each point in the graph represents an average of 100 replications of the experiment, each run for 10 000 customers.

As seen in the graph, the set of R3 policies reduces the mean system delay considerably, especially at heavy load where it pays off significantly to avoid routing consecutively to the same queue. It is worth noting that the set of semi-R3 policies does indeed perform better than the set of R3 policies at light traffic (although the difference in the mean delay is small, but observable within the confidence interval of our simulation, and is not clearly seen in the graph). This was first noted in [13] in the context of queues with deterministic servers. The intuition lies on a trade-off between the degree of sharing of individual servers by customers from different arrival streams and the variability in the arrival process that is seen by a particular server.

IV. ROUTING IN HETEROGENEOUS SYSTEMS

In this section, we consider a system consisting of K heterogeneous queues each of which has infinite capacity and a dedicated server. The service times at queue i form a sequence of i.i.d. exponential random variables with rate μ_i . There is one arrival stream (see Fig. 3) and a controller that has no queue length information and must select a queue for each customer that arrives only based on information regarding its past routing decisions. We assume that the arrival process is independent from all the service processes but is otherwise arbitrary. Let Σ_h denote the class of feasible routing policies.

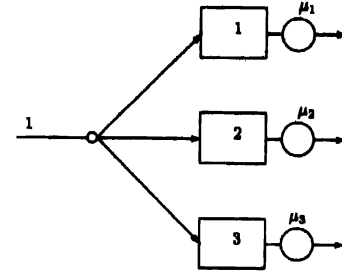


Fig. 3. A system with heterogeneous servers.

We use the term probabilistic policy to refer to any routing policy in Σ_h that uses Bernoulli splitting, i.e., routes an arriving customer to queue i with probability p_i , $\sum_{i=1}^K p_i = 1$, regardless of past decisions. Our objective is to show that, given any probabilistic routing policy π , there exists another policy ρ which partly employs an R3 algorithm, and which performs better than π in the sense of separable increasing convex ordering of the queue lengths. In particular, policy ρ uses a Bernoulli algorithm to split the arrival process into a finite number of subprocesses, effectively creating a finite number of substreams of customers. Then, for each substream, ρ either routes all customers to a particular queue or applies an R3 algorithm, possibly only to a subset of the K queues. The exact way in which the splitting of the arrival process is done depends on the routing probabilities for π and is described in the following. We prove our result for $K = 2$ and later discuss how the result can be generalized to arbitrary K .

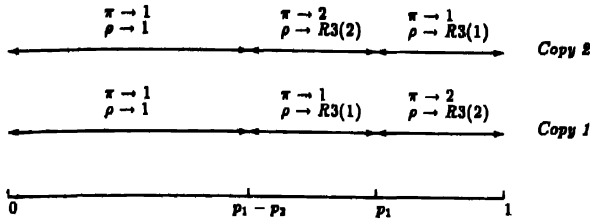
Theorem 1: For any probabilistic policy π in Σ_h , employed in a system with two queues, there exists a policy ρ in Σ_h that partly employs an R3 algorithm such that the following is true:

$$\{N^\rho(t); t \geq 0\} \leq_{I_1} \{N^\pi(t); t \geq 0\}$$

provided that $N_i^\rho(0) =_{st} N_i^\pi(0)$ for all $i, j = 1, \dots, K$.

Proof We condition on initial queue lengths, arrival times, and service times. We assume that service completions at queue i can only occur at the jump points of a Poisson process with rate μ_i . Consider a probabilistic policy π that routes an arriving customer to queue 1 or 2 with probabilities p_1, p_2 , respectively, $p_1 + p_2 = 1$. Without loss of generality, we can assume that $p_1 \geq p_2$. The proof is by construction. Specifically, we construct a policy ρ as follows. When a customer arrives, ρ draws a random number u uniformly from the interval $[0, 1]$. If $u \in [0, p_1 - p_2]$, the customer is routed to queue 1. Otherwise, the customer is routed to either queue 1 or queue 2 according to an R3 algorithm applied to queues 1 and 2. That is, for the subsequence of arrivals associated with random numbers that fall in the interval $(p_1 - p_2, 1]$ policy ρ uses an R3 routing policy.

For each of the two policies, i.e., π and ρ , we create two copies of the system, as follows. First, we divide the interval $I = [0, 1]$ into three subintervals, namely, $I_1 = [0, p_1 - p_2]$, $I_2 = (p_1 - p_2, p_1]$, and $I_3 = (p_1, 1]$. In the first copy of the system, π draws a random number u from $[0, 1]$ and routes to queue 1 if $u \in I_1 \cup I_2$, otherwise it routes to queue 2. Policy ρ , on the other hand, draws a random number u from $[0, 1]$ and routes to queue 1 if $u \in I_1$, otherwise uses an R3 algorithm that starts from queue 1 if the first random number u that falls in $I_2 \cup I_3$ is such that $u \in I_2$; otherwise, it starts the algorithm from queue 2. In the second copy of the system, π routes to queue 1 if $u \in I_1 \cup I_3$, otherwise it routes to queue 2. Policy ρ , routes to queue 1 if $u \in I_1$, otherwise uses an R3 algorithm that starts from queue 1 if the first random number u that falls in $I_2 \cup I_3$ is such that $u \in I_3$; otherwise, it starts the algorithm from queue 2 (see Fig. 4).


 Fig. 4 The construction $R3(i)$ denotes $R3$ starts from queue i

To distinguish between the two copies we use the notation $(n_1^1(t), n_2^1(t))$ and $(m_1(t), m_2(t))$ to denote the vectors of the queue lengths for the first and the second copy respectively. $\pi = \rho \circ \pi$. Noting that $|I_1 \cup I_2| = |I_1 \cup I_1| = p_1$ and $|I_2| = |I_1| - p_2$ it is seen that the marginal queue lengths in each of the two copies are stochastically equal to those of the original system for either π or ρ . Coupling the random numbers that ρ and π use for making routing decisions, the key of the proof is in showing that the following relations are true on a sample path for all $t > 0$:

$$(n_1^1(t), m_1^1(t)) \leq (n_1^\pi(t), m_1^\pi(t)) \quad (1)$$

$$(n_2^1(t), m_2^1(t)) \leq (n_2^\pi(t), m_2^\pi(t)) \quad (2)$$

We only discuss (1); relation (2) follows similarly. Consider the two cross systems namely $(n_1^1(t), m_1^1(t))$ and $(n_1^\pi(t), m_1^\pi(t))$. By assumption, service completions occur simultaneously in these two systems (at the jump points of a Poisson process with rate μ_1) and so do arrivals associated with random numbers drawn from I_1 . As to arrivals associated with random numbers from $I_2 \cup I_1$, note that under ρ these alternate between the queues with lengths $n_1^1(t)$ and $m_1^1(t)$. Thus the cross system under ρ (see the left hand side of (1)) can be thought as one consisting of two queues with homogeneous exponential servers, two dedicated arrival streams (arrivals associated with us in I_1) and one common external stream (arrivals associated with us in $I_2 \cup I_1$) with a controller that uses an RR policy. The same is true for the cross system under π except that the controller on the external stream does not use an RR policy. Thus (1) follows by the arguments preceding Proposition 1.

Let $\mathbf{n}^\pi(t) = (n_1^\pi(t), n_2^\pi(t), m_1^\pi(t), m_2^\pi(t)) = \rho \circ \pi$. From (1) (2) and an elementary property of majorization it follows that

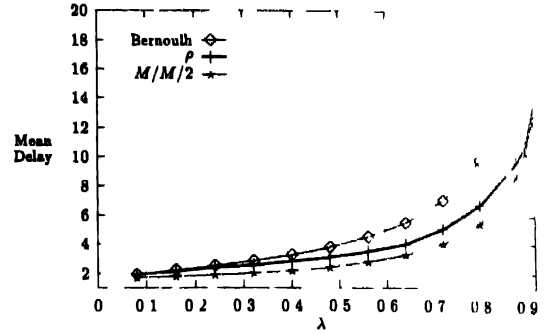
$$(\mathbf{n}^1(t_1), \mathbf{n}^1(t_2)) \leq (\mathbf{n}^\pi(t_1), \mathbf{n}^\pi(t_2)) \quad \forall t_1, t_2$$

Thus

$$\begin{aligned} \sum_{i=1}^l [g(n_1^1(t_i)) + g(n_2^1(t_i))] + \sum_{i=1}^l [g(m_1^1(t_i)) + g(m_2^1(t_i))] \\ \geq \sum_{i=1}^l [g(n_1^\pi(t_i)) + g(n_2^\pi(t_i))] \\ + \sum_{i=1}^l [g(m_1^\pi(t_i)) + g(m_2^\pi(t_i))] \quad (3) \end{aligned}$$

for all g increasing and convex. Taking expectations and recalling that the marginal queue lengths for each of the two copies are stochastically equal to those of the original system for both ρ and π the result follows. ■

As it is seen from the proof, ρ can be easily implemented. Let us now describe how ρ is specified for arbitrary K . For simplicity assume the routing probabilities for π are in decreasing order, i.e. $p_1 \leq p_2 \leq \dots \leq p_K$. Then ρ acts as follows: 1) uses an R3 algorithm applied to queues 1 to K for a Kp_1 portion of the incoming traffic; 2) uses another independent R3 algorithm applied


 Fig. 5 Comparing π (Bernoulli) and ρ

to queues 2 to K for a $(K-1)(p_1 - p_1)$ portion of the incoming traffic. $(K-1)$ uses an R3 algorithm applied to queues $K-1$ and K for a $2(p_{K-1} - p_K)$ portion of the incoming traffic and finally (K) routes all customers from the remaining $(p_K - p_{K-1})$ portion of the traffic to queue K . Note that these portions of the incoming traffic are specified by a probabilistic routing algorithm.

In the interest of space we will sketch the proof for $K=3$. This essentially describes the inductive argument that extends the proof to arbitrary K . First define a policy ρ_1 as follows. Policy ρ_1 uses an R3 algorithm applied to queues 1, 2, 3 for a $3p_1$ portion of the traffic; routes a $2(p_1 - p_1)$ portion of the traffic to queue 2 and a $p_1 - p_2$ portion of the traffic to queue 3. To prove that ρ_1 is better than π one needs to construct three copies of the system such that the order of arrivals to queues 1, 2, 3 under R3 is 1, 2 and 3 for copy 1, 2, 3 and 1 for copy 2 and finally 3, 1 and 2 for copy 3. Then similarly to Theorem 1 (see (1) and (2)) it can be shown that

$$(n_1^1(t), n_2^1(t), n_3^1(t)) \leq (n_1^\pi(t), n_2^\pi(t), n_3^\pi(t)) \quad i=1, 2, 3 \quad (4)$$

where n_i denotes the queue length process at queue i in copy j under policy ρ_1 .

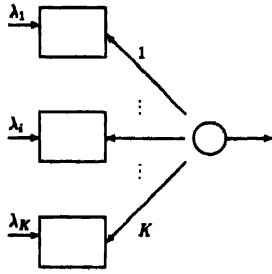
Now let ρ be a policy that uses an R3 algorithm applied to queues 1, 2, 3 for a $3p_1$ portion of the traffic, uses another R3 algorithm applied to queues 2, 3 for a $2(p_1 - p_1)$ portion of the traffic and routes a $p_1 - p_2$ portion of the traffic to queue 3. An application of Theorem 1 to queues 2 and 3 proves that ρ outperforms ρ_1 . Thus ρ is the optimal policy in Σ_1 .

A numerical example is given in Fig. 5. There are two queues with service times that are exponential, the rates are $\mu_1 = 0.6$, $\mu_2 = 0.1$. Customers arrive from a Poisson stream of rate λ . The optimal routing probabilities for a Bernoulli policy (in terms of minimizing the steady state mean system delay) are known (e.g. [18]) to be given by

$$p_1 = \min \left\{ \frac{\mu_1 \sqrt{\mu_2} - \mu_2 \sqrt{\mu_1} + \lambda \sqrt{\mu_1}}{\lambda(\sqrt{\mu_1} + \sqrt{\mu_2})}, 1 \right\} \quad p_2 = 1 - p_1 \quad \mu_1 \geq \mu_2 \quad (5)$$

The graph compares the mean delay under a probabilistic routing policy π that uses the aforementioned routing probabilities with a policy ρ that routes a $(p_1 - p_2)$ portion of the incoming traffic to queue 1 and uses an R3 algorithm for the remaining portion of the traffic. The graph also presents a lower bound on the performance of policies that use no queue length information, expressed as the mean delay of a preemptive M/M/2 queue, that is the slow server, i.e., the one with rate μ_2 , can be preempted so that the fast server is always kept busy as long as there are customers in the system. This bound will, of course, never be achieved.

As it is seen in the graph, policy ρ reduces the mean system delay considerably at heavy load, compared to π . At light load all

Fig. 6. A scheduling system with K queues.

three policies behave approximately the same, as the fast server is almost exclusively used. In particular, the Bernoulli policy and ρ have exactly the same performance since under the former, customers are always routed to the fast server: for $\lambda < 0.1$, it follows from (5) that $p_1 = 1$. Thus, by the way ρ is constructed, it follows that it also always routes to the fast server. On the other hand, at heavy load, the performance of ρ approaches that of the $M/M/2$ system. In contrast, the Bernoulli policy increases the mean system delay away from its lower bound.

V. THE SCHEDULING PROBLEM

In this section, we consider a system consisting of K queues each of which has infinite capacity. Queue i receives customers from a Poisson stream with rate λ_i . There is a single server that may switch to one of the queues according to a scheduling policy that is determined by a controller (see Fig. 6). We assume that the controller has no queue length information and schedules the server based only on the history of its past decisions. Once the controller selects a queue to which the server is scheduled, we assume that a single customer from that queue is served (if one exists) before the server switches to another queue. If immediately upon the time the server joins a queue, that queue is empty, it is assumed that the server remains idle for a time period equal to the service time of a single customer; then it joins another queue. This is a reasonable assumption, for example, when time is slotted and all service times are deterministic equal to the duration of a time slot. In the following, we assume that the distribution of the service times is general and that the arrival processes are mutually independent as well as independent of the service processes.

The interest in studying these systems primarily arises in the context of multiple access communication systems. One may think of K stations sharing a common channel and wishing to transmit fixed-length packets at the beginning of time slots. One of the types of policies that can be used for accessing the channel are collision-free policies, under which it is not possible for two or more stations to attempt transmission of a packet at the beginning of the same time slot. Such policies can be deterministic, e.g., time division multiplexing, or probabilistic. The latter can be implemented in a distributed fashion, i.e., without any need for centralized coordination, provided that all stations use the same random number sequence.

Let $\Sigma_{s,h}$ be the class of scheduling policies for the system that we have just described. The result in this section is similar to that proved in Section IV. That is, given any probabilistic scheduling policy π in $\Sigma_{s,h}$, there exists another policy ρ that partly employs an R3 algorithm, and that outperforms π in the sense of separable increasing convex ordering of queue lengths. Then, ρ can be easily constructed on basis of these probabilities and be guaranteed to reduce the total workload at all times. The proof and the construction are similar to those in Section IV and are omitted.

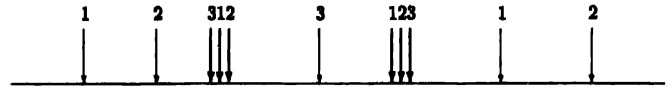


Fig. 7. Routing customers to queues. Arrivals from the external stream (thick lines) are as if the RR algorithm applied only to them.

Theorem 2: For any probabilistic policy π in $\Sigma_{s,h}$, there exists a policy ρ that partly employs an R3 algorithm such that the following is true:

$$\{N^{\rho}(t); t \geq 0\} \leq_t \{N^{\pi}(t); t \geq 0\}$$

provided that $N_i^{\rho}(0) =_{st} N_j^{\pi}(0)$ for all $i, j = 1, \dots, K$.

Remark 2: It is not difficult to see that the arrival processes need not, in fact, be Poisson. Note that in the routing problem, we were not able to relax the assumption of the service times being exponentially distributed, since we required that both real and fictitious service completions be synchronized at the jump points of the service-counting process. This simplifies the proof considerably. In the scheduling system, of course, arrivals are always real, thus, there is no need to use the memoryless property of the exponential distribution.

VI. EXTENSIONS

This section contains a discussion of further issues that arise in the context of routing. Similar remarks can be made, of course, for the scheduling problem. First, it is interesting to study systems with finite capacities. A slight extension of Theorem 1 in [12] shows that in a system with M external arrival streams and K homogeneous exponential servers, a set of R3 policies minimizes the expected number of departures by any time t and, at the same time, it minimizes the expected number of losses by t . As a consequence, Theorem 1 can be extended to systems with finite buffers. As one expects, due to losses that may occur, the E_t^1 ordering on the queue lengths no longer holds when capacities are finite.

Theorem 1 does in fact hold for IHR service-time distributions. The key is in extending Theorem 1 in [6] (that proves the optimality of the RR policy when the service times form a sequence of i.i.d. r.v.s with IHR distribution) to systems that may admit arrivals from K dedicated arrival streams. The extension is rather straightforward. Since the arrival process in [6] is assumed to be arbitrary, one can simply assume that arrivals from the dedicated streams that can be assumed to occur simultaneously at the K queues are part of the external arrival process and are routed according to the RR policy. This does not modify the algorithm for the original common external arrival process (see, for example, Fig. 7, where $K = 3$).

ACKNOWLEDGMENT

The author is grateful to Prof. C. Cassandras and Prof. D. Towsley for their support, encouragement, and invaluable advice. The author is also thankful for many discussions with them that stimulated this research and improved the final version of this note.

REFERENCES

- [1] Y. C. Chow and W. H. Kohler, "Models for dynamic load-balancing in a heterogeneous system," *IEEE Trans. Computers*, vol. 28, pp. 354-361, 1979.
- [2] R. L. Cruz and M. C. Chuah, "A minimax approach to a simple routing problem," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1424-1435, 1991.
- [3] A. Ephremides, P. Varaiya, and J. Walrand, "A simple dynamic routing problem," *IEEE Trans. Automat. Contr.*, vol. 25, 1980.
- [4] A. Ephremides and T. V. Truong, "Scheduling broadcasts in multihop radio networks," *IEEE Trans. Commun.*, vol. 38, no. 4, 1990.

- [5] A. Hordijk and G. Koole, 'On the optimality of the generalized shortest queue policy,' *Probability in the Eng. Info. Sci.*, vol. 4, pp. 477-487, 1990.
- [6] Z. Liu and D. Towsley, 'Optimality of the round robin routing policy,' *COINS Tech. Rep. 92-55*, Univ. Mass., 1992.
- [7] —, 'Effects of service disciplines in G/G/s queueing systems,' submitted for publication, 1992.
- [8] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. New York: Academic, 1979.
- [9] R. Menich and R. F. Serfozo, 'Optimality of routing and servicing in dependent parallel processing systems,' *Queueing Syst.*, vol. 9, pp. 403-418, 1991.
- [10] Z. Rosberg and D. Towsley, 'Customer routing to parallel servers with different rates,' *IEEE Trans. Automat. Contr.*, vol. 30, pp. 1140-1143, 1985.
- [11] S. M. Ross, *Stochastic Processes*. New York: Wiley, 1983.
- [12] P. D. Sparaggis, D. Towsley, and C. G. Cassandras, 'Optimality of static routing policies in queueing systems with blocking,' *Proc. 1991 IEEE Conf. Decision Contr.*, Brighton, England, 1991, pp. 809-814.
- [13] G. D. Stamoulis and J. N. Tsitsiklis, 'Optimal distributed policies for choosing among multiple servers,' *Proc. 1991 IEEE Conf. Decision Contr.*, Brighton, England, 1991, pp. 815-820.
- [14] A. Tantawi and D. Towsley, 'Optimal static load balancing in distributed computer systems,' *J. Assoc. Comput. Mach.*, vol. 32, pp. 445-465, 1985.
- [15] L. Tassiulas and A. Ephremides, 'Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks,' *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1936-1948, 1992.
- [16] D. Towsley, P. D. Sparaggis, and C. G. Cassandras, 'Optimal routing and buffer allocation for a class of finite capacity queueing systems,' *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1446-1451, 1992.
- [17] D. Towsley and P. D. Sparaggis, 'Optimal routing in systems with II-R service time distributions,' submitted for publication, 1993.
- [18] J. Walrand, *An Introduction to Queueing Networks*. Englewood Cliffs, NJ: Prentice Hall, 1988.
- [19] R. R. Weber, 'On the optimal assignment of customers to parallel queues,' *J. Appl. Prob.*, vol. 15, pp. 406-413, 1978.
- [20] W. Whitt, 'Deciding which queue to join,' *Operations Res.*, vol. 34, pp. 55-62, 1986.
- [21] W. Winston, 'Optimality of the shortest line discipline,' *J. Appl. Prob.*, vol. 14, pp. 181-189, 1977.

Methods and Theory for Off-Line Machine Learning

S. Yakowitz and J. M. Wu

Abstract—Many problems in machine learning can be abstracted to the sequential design task of finding the minimum of an unknown, often and possibly discontinuous function on the basis of noisy measurement. In the present work, it is presumed that there is no penalty for bad choices during the experimental stage, and at some time, not known to the decision maker, or under his control, the experimentation will be terminated, and the decision maker will need to specify the point considered best, on the basis of the experimentation. In this note, we seek the best trade-off between i) acquiring new test points, and ii) retesting at points previously selected so as to improve the estimates of relative performance. The algorithm is shown to achieve a performance standard described herein. This decision setting would seem natural for function minimization in a simulation context or for tuning up a production process prior to putting it into service.

1. INTRODUCTION

There are myriads of situations in which one is confronted with a sequence of stochastically similar decision problems. In such situations, as in real life, one often has the opportunity to select better decisions as a data base of actions and outcomes accumulates. When this sifting and exploration of decision space is done by a computer algorithm, we term the process machine learning. Such an activity can be abstractly viewed as the process of finding the minimum of an unknown function on the basis of noisy measurements. In the present note, the focus is on this abstraction.

A characteristic of off-line learning is that there is no need to be concerned with performance during the exploration stage. (By contrast, on-line learning seeks to minimize observed loss as the process evolves.) The criterion of 'goodness' of off-line algorithms is that the decision or design provided at the end of the session performs as well as possible. It is tacitly assumed that this decision point will be selected for implementation without further testing. For example, in blackjack, during simulations, no money is to be gained or lost. In a long sequence of exploratory games, one could occasionally test to see how good the best tens count strategy found thus far is, and end the experimental session and head off to Nevada when performance seems promising. The blackjack study [1], however, employed on-line learning, but in view of results to follow, it appears that the learning plan could have been sharpened further.

The goal of the present theoretical study is to pose and examine off-line learning algorithms. We are aware of a lively literature (e.g., Devroye [2], [3], Jarvis [4], Guin [5], and Matyas [6]) which includes convergence results, but to our knowledge, rates have not been established for off-line methods. The conclusion substantiated herein is that significantly faster convergence rates can be achieved off-line under various hypotheses than by on-line algorithms.

We proceed now with the particulars. A formal statement of the problem we intend to address is as follows:

Definition 1.1—Let $f(\cdot)$ denote a bounded real valued measurable function on an infinite Borel action set $\mathcal{A} \subset R^1$. The off-line learning problem is the task of sequentially

Manuscript received November 17, 1992; revised May 21, 1993; December 10, 1993; and March 21, 1994. The work of S. Yakowitz was supported in part by the National Science Foundation under grants ECS 8913642 and JNT 9201430 and by the National Institutes of Health under grant A129426-01.

The authors are with the Systems and Industrial Engineering Department, University of Arizona, Tucson, AZ 85721, USA.

IEEE Log Number 9407033.

1) choosing design points x_n , on the basis of a history of noisy observations

$$y_j = y(x_j) = f(x_j) + \epsilon(x_j), \quad j < n, \quad (1)$$

and

2) designating a particular design point x_n^* on the basis of these observations, so that for some specified positive number ϵ , as $n \uparrow \infty$

$$P[f(x_n^*) - f_{\min} > \epsilon] \rightarrow 0. \quad (2)$$

Convergence rate assurances are desirable.

In (2) f_{\min} is the minimum (or Lebesgue-essential infimum) of $f(x)$. Also $\epsilon(x_j)$ in (1) is a 0-mean random variable and the sequence $\{\epsilon(x_j)\}$ is hypothesized to satisfy certain assumptions to be stated in Section II.

One can intuitively justify an off-line learning criterion by imagining that one can perform experiments until some unknown time when it will be demanded that the experimentation cease and the best operating point, i.e., x_n^* , found so far be declared and used.

II. AN OFF-LINE LEARNING ALGORITHM

A. The Off-Line Method

The simple plan of the algorithm is occasionally to add a new randomly chosen design point to our list of test points. At other times, one refines the estimate of the expected performance of design points already on the list.

We first offer a generic off-line learning algorithm, and the subsequent discussions will be aimed at stating properties of this algorithm under various postulates. In what follows, the symbol \mathcal{C} will represent various positive constants, the values of which are not important for the associated discussions. The algorithm has as parameters a probability density function $p(\cdot)$ on the action set \mathcal{X} and a nondecreasing integer sequence $\{NP(n)\}_n$. The density does not affect the rate of convergence (provided it satisfies postulates to be stated). If the user has some intuition about the location of promising points in decision space \mathcal{X} , this could be reflected by relatively large density values. The $NP(n)$ sequence specifies times at which new decision points should be acquired and added to our experimentation list. The selection of this sequence should depend on properties one is willing to assume about the noise. In Section II-C, the important cases of finite moment and finite moment generating function are covered.

Algorithm Parameters: i) $p(x)$ is a probability density function having as support all of a (measurable) decision space $\mathcal{X} \subset \mathcal{R}^d$. (If \mathcal{X} is discrete, take $p(x)$ to be a probability mass function.)

ii) $\{NP(n)\}$ is a nondecreasing unbounded sequence of integers such that

$$NP(1) = 1, \quad NP(n)/n \downarrow 0, \quad \text{and} \quad \forall P(n) - NP(n-1) \in \{0, 1\}$$

If $\sup_x E|\epsilon(x)|^p < \infty$, let $NP(n) \propto \ln(n)$.

If $\sup_x E|\exp(\theta\epsilon(x))| < \infty$, let $NP(n) \propto \sqrt{n}$.

Set $n = 1$ and proceed to the iterative step.

The Iterative Step:

New Sample Times (Steps 1 and 2): If $n = 1$ or $NP(n-1) < NP(n)$, randomly select and test a new point from \mathcal{X} .

1) Choose a point, designated by $t(NP(n))$, from \mathcal{X} at random according to the pdf $p(x)$. Set $x_n = t(NP(n))$ and observe $y_n = y(x_n)$.

2) Start a running average $m_{NP(n)}$ for observations at $t(NP(n))$ by declaring

$$m_{NP(n)} = y_n$$

and start a counter at $t(NP(n))$ by setting $NS(NP(n)) = 1$.

Remark: The user-specified sequence " $NP(n)$ " (See item ii)) tells the number of New [sample] Points that have been gathered by time n . The random quantity " $NS(j)$," which is determined by the realization of the learning process, gives the number of samples that have been taken at a given sample point $t(j)$, $j \leq NP(n)$.

Go to 6.

Resample Times (Steps 3, 4, and 5): Else if $NP(n-1) = NP(n)$, resample at the point which has been tested the fewest number of times.

3) Let I be any index i , $1 \leq i \leq NP(n)$, such that

$$NS(i) \leq NS(j), \quad 1 \leq j \leq NP(n). \quad (3)$$

Set $x_n = t(I)$, and observe y_n .

4) Update the sample mean and sample counter. Set

$$m_I = [m_I \cdot NS(I) + y_n] / (NS(I) + 1). \quad (4)$$

and

$$NS(I) = NS(I) + 1. \quad (5)$$

5) At all times n such that $\forall P(n) > 1$, announce as the best point (BP) the test point $t(I^*)$, with $I^* = i$ where i is any index less than $NP(n)$ such that

$$m_i \leq m_j, \quad 1 \leq j < NP(n). \quad (6)$$

(Note that $t(NP(n))$ is not a candidate for the BP because it has recently been selected, $NS(NP(n)) \ll NS(j)$, $j < NP(n)$. Thus, the precision of the running mean $m_{NP(n)}$ as an estimate of $f(x_{NP(n)})$ is worse than that of the precision of the other running means.)

6) Set $n = n + 1$. Repeat the iterative step. (There is no official stopping condition, although one anticipates that at some arbitrary time, not known to the designer, the experimentation will be halted.)

[End of Algorithm]

The properties of the algorithm will require various hypotheses, some of which will always be assumed in force. These are now presented.

B. Convergence Analysis

Action Space Assumption: The set \mathcal{X} is either discrete (but infinite) or has positive Lebesgue measure.

Noise Hypothesis 1: For any $t \in \mathcal{X}$, define $I_t = \{i(x_j) : x_j = t\}$. Then this random subsequence, if not empty, is independent and identically distributed (i.i.d.). Any collection $I_{t(1)}, I_{t(2)}, \dots$ of such sequences are mutually stochastically independent.

Noise Hypothesis 2: Let us suppose that for any error tolerance d and any number N of observations of $Y(x)$, at any fixed x , there is a known function $P(d, N)$ which satisfies the following conditions:

i) uniformly in $x \in \mathcal{X}$

$$P[|m_{\cdot}(N) - f(x)| > d] \leq P(d, N). \quad (7)$$

Here, we have let $m_{\cdot}(N)$ represent the mean of a sample of $Y(x)$ of size N at design point x .

ii) If N is regarded as a variable on the positive numbers, then for any d , $P(d, N)$ is continuously differentiable and strictly convex.

The off-line algorithm assures that the number of observations, $NS(i)$, made at each test point $t(i)$, is about the same for each i , $1 \leq i < NP$. A property of this strategy which is fundamental to our analysis is the following proposition.

Proposition 2.1 Assume property ii) of the Noise Hypothesis 2. Then among all static algorithms for allocating M observations among a given number $\backslash P$ of test points the allocation $\Lambda \propto M/\backslash P$ to each test point $t(i)$ is the one which minimizes

$$\sum_{i=1}^{\backslash P} P(d \backslash \Lambda_i) \quad (8)$$

Proof By the Lagrange multiplier procedure, the minimizer of (8), subject to $\sum_{i=1}^{\backslash P} \Lambda_i = M$ must satisfy for each i , $1 \leq i \leq \backslash P$ and for some number λ

$$\partial/\partial \Lambda_i \left\{ \sum_{i=1}^{\backslash P} P(d \backslash \Lambda_i) + \lambda \left(\sum_{i=1}^{\backslash P} \Lambda_i - M \right) \right\} = 0 \quad (9)$$

from which we have the following necessary condition

$$\partial/\partial \Lambda_i P(d \backslash \Lambda_i) = -\lambda \quad 1 \leq i \leq \backslash P \quad (10)$$

But this evidently implies that the derivatives of $P(d \backslash \Lambda_i)$ must all be equal. By the strict convexity hypothesis the derivatives are strictly monotonically increasing so the value at which they equal any given value $-\lambda$ must be unique. That in turn implies uniqueness of the optimizing value Λ_i . That is all the Λ_i s are the same. QED

In this study the objective of learning is to try to assure that at the designated best design point the objective function value is within some given tolerance ϵ of optimality. That is in reference to Algorithm Step 5 we seek a testing plan which guarantees that

$$P[\text{Mistake}(n)] = P[f(t(I^*)) > f_{\text{opt}} + \epsilon] \downarrow 0 \quad (11)$$

at some assured rate. Let

$$f_{\text{opt}}(n) = \min_{t \in T(n)} f(t(i))$$

We construct a sequence giving an assured rate by noting that for functions $Q_1(\backslash P(n)/n)$ and $Q_2(\backslash P(n)/n)$ defined below

$$\begin{aligned} P(\text{Mistake}(n)) &\leq P[f_{\text{opt}}(n) > f_{\text{opt}} + \epsilon/2] \\ &\quad + P\left[\max_i |m_i - f(t(i))| > \epsilon/2\right] \\ &\leq Q_1(\backslash P(n)/n) + Q_2(\backslash P(n)/n) \end{aligned} \quad (12)$$

It will be apparent that for fixed n as $\backslash P$ increases Q_1 decreases while Q_2 gets larger. So the logical thing to do is to balance these probabilities. Specifically select $\backslash P^*(n)$ to assure that for some constants $C_1 > C_2 > 0$ and all n

$$C_1 Q_1(\backslash P^*(n)/n) \leq Q_2(\backslash P^*(n)/n) \leq C_2 Q_1(\backslash P^*(n)/n) \quad (13)$$

Toward achieving this design define $F(d) = P[f(T) < d]$ where T is distributed according to the algorithm search density $p(\cdot)$. Then $Q_1(\backslash P(n)/n) = (1 - F(\epsilon/2))^{1/\backslash P(n)}$. Under the independence assumption of Noise Hypothesis 1

$$\begin{aligned} P\left[\max_i |m_i - f(t(i))| > \epsilon/2\right] \\ \leq Q_2(\backslash P(n)/n) \\ = \backslash P^*(n) P(\epsilon/2 \backslash n/\backslash P^*(n)) \\ (1 + q(\backslash P^*(n) P(\epsilon/2 \backslash n/\backslash P^*(n)))) \end{aligned}$$

where $q(\backslash P^*(n) P(\epsilon/2 \backslash n/\backslash P^*(n))) = o(\backslash P^*(n) P(\epsilon/2 \backslash n/\backslash P^*(n)))$. Thus by simply equating Q_1 and Q_2 we obtain the implicit relation

$$\backslash P^*(n) \sim \ln(\backslash P^*(n)) P(\epsilon/2 \backslash n/\backslash P^*(n)) / \ln(1 - F(\epsilon/2)) \quad (14)$$

Since $\ln(\backslash P^*(n))/\backslash P^*(n) \rightarrow 0$ it suffices for $t(i)$ to satisfy the relation

$$\backslash P^*(n) = \text{Integer Part} [P(\epsilon/2 \backslash n/\backslash P^*(n)) / \ln(1 - F(\epsilon/2))]$$

We summarize the developments on an achievable rate in Proposition 2.2

Proposition 2.2 Under the preceding noise and decision hypotheses if $\backslash P^*(n)$ satisfies (15) then given any tolerance ϵ for some $0 < \rho < 1$

$$P[f(t(I^*)) > f_{\text{opt}} + \epsilon] = O(\rho^{n^{1/2}}) \quad (16)$$

Usually the value $F(\epsilon/2)$ is not known to the designer but (16) remains satisfied for any positive value $\rho < 1$ of this constant. The effect of omitting the divisor in (15) is perhaps to require a value ρ closer to 1.

C Applications and Implications

The cases of obvious interest are when the noises are known to have uniformly bounded moments of some order $q > 1$ or alternatively are known to have a moment generating function. Thus in the notation of (1), respectively

$$\sup |F(t(i))| \leq C \quad (17)$$

or

$$\sup E[\exp(\theta t(i))] \leq C \quad (18)$$

for all θ in some interval I . Between the two conditions of finitely many moments (17) and infinitely many moments (18) essentially all of the well known random variable families will be accounted for in the following statement

Proposition 2.3 Under the conditions of the Proposition 2.2 for the moment condition (17) the condition (13) is satisfied if

$$\backslash P(n) \sim C \log(n) \quad (19)$$

and for (18) if

$$\backslash P(n) \sim C \sqrt{n} \quad (20)$$

In the moment bound case we have

$$P[\text{Mistake}(n)] = O(n^{-(1/2)} \log(n)) \quad (21)$$

and in the finite moment generating function setting

$$P[\text{Mistake}(n)] = O(\rho^{n^{1/2}}) \quad (22)$$

for some (ϵ dependent) number ρ in the open unit interval

Proof For (19) a moment/probability bound e.g. [7] under the moment condition (17) is that

$$P(d \backslash n) \leq C/n^{1/2}$$

Equation (21) results from substitution of this bound into (15). Then by the equation preceding (14) and the balance relation (13) we have

$$P[\text{Mistake}(n)] \leq C \log(n)/n^{1/2} \quad (23)$$

which gives (21)

Similarly from Cramer's theorem e.g. [8 p. 9] we have the assurance that the existence of the moment generating function implies the exponential inequality for (7)

$$P(d \backslash n) = C \exp(-Qn) \quad (24)$$

where $Q = Q(d)$ is the dominating point of the large deviation rate function. See [8] for details on its calculation. In the Gaussian case

$Q(d) = d^2/2\sigma^2$ where σ^2 is an upper bound for the variances of $\epsilon(x)$, $x \in \mathcal{X}$.

Set $\rho < 1$ to be some number slightly larger than $\exp(-Q(d))$, say $\rho = J \exp(-Q(d))$ and note that for any $j > 1$, $j^n > \sqrt{n}$ for n sufficiently large. In view of (24), this gives (20), and (22) by the arguments for the moment case. QED

Under the circumstance that either the noise has a symmetric distribution, or that the distribution function of the noise is given, one can estimate $f(t(i)) + C$, for some constant C not depending on $t(i)$, by med_i , the sample median at $t(i)$. In the known-noise case, if the median is not defined, use the sample minimum of any defined quantile. For quantiles, probability bounds of type (20) are satisfied [9]. That is, for fixed $d > 0$, and some positive constant C'

$$P[|\text{med}_i - f(t(i)) + C| > d] = O(\exp(-C' NS(i))), 1 \leq i \leq NP.$$

The analysis for the moment generating function case then implies that for $t(I^*)$ the minimizer of the sample medians, and for $NP(n) \propto \sqrt{n}$, the probability of mistake is $O(\rho^{\sqrt{n}})$, some $0 < \rho < 1$. These observations might be useful if the noise has undefined expectation, as in the case of the Cauchy family.

We now argue that the rate cannot be uniformly improved by any other sampling plan.

Proposition 2.4: Over the class of noise processes satisfying (18), the fastest achievable rate of convergence of probability of mistakes to 0, with increasing decision time n , is $O(\rho^{\sqrt{n}})$, for some (problem dependent) number ρ in the open-unit interval.

Proof: It suffices to exhibit a single learning problem for which no improvement over $O(\rho^{\sqrt{n}})$ is possible. The waiting time until a sample point T satisfies $f(T) < f_{\min} + \epsilon/2$ is geometrically distributed. The only conceivable way to improve upon the conjectured rate is to gather new design points at a rate faster than \sqrt{n} . This, of course, entails requiring that $NS(i)/\sqrt{n} = o(1)$ at all design points $t(i)$: $i < NP(n)$.

Now suppose that unknown to the decision maker, $f(x)$ only has two points, say 0 and 1, in its range, and that $0 < p_1 := P(f(T) = 1) < 1$. Further, it is presumed that the noise $\epsilon(t) = 0$ if $f(t) = 0$ and standard normal otherwise, and that the tolerance $\epsilon = 1/2$, so that a mistake occurs whenever $f(t(I^*)) = 1$. It is now easy to confirm that for "bad" indexing the sample points $t(j)$: $f(t(j)) = 1$, and m_j being the sample mean or median at t_j , that for some number $C' > 0$

$$P\left[\min_{j \in \text{bad}} m_j < 0\right] \geq \exp(-C' \cdot NSM(n)) \quad (25)$$

where $NSM(n) = \min NS(i)$, $i < NP(n)$. Now from (25) it is clear that the probability that $f(t(I^*)) = 1$ cannot diminish faster than $\rho^{NSM(n)}$ with $\rho = \exp(-C')$. Thus, any sampling plan for which $NSM(n) = o(\sqrt{n})$ will have larger asymptotic probability of mistakes than $NP^*(n)$ in (15).

D. Comparison to On-Line Learning Rates

The on-line criterion is that for x_n the point tested at epoch n , a mistake occurs if

$$f(x_n) - f_{\min} > \epsilon.$$

Yakowitz and Lowe [10] have analyzed the expected number of mistakes in the first n decision times. The expected number of mistakes can be written as

$$J(n) = \sum_{i=1}^n P[f(x_i) > f_{\min} + \epsilon]. \quad (26)$$

This criterion can be viewed as the measure of cumulative performance in the control process, up to time n . Since it measures the cost

of each control x , actually applied, as opposed to what is estimated to be the best control found so far, i.e., $t(I^*)$, there is the potential that the expected performance improvement is slower than for off-line learning.

Toward comparison with Proposition 2.3, one can approximate the probability $P_{ON}(n)$ of error in the on-line case by the first difference,

$$P_{ON}(n) = \nabla J(n) = J(n+1) - J(n) \sim P[\text{on-line Mistake}(n)].$$

From Theorem 3 of the aforementioned work, for the moment bound case (17) and the exponential case (18), it turns out that the corresponding rates of convergence to 0 are, respectively,

$$P_{ON}(n) = O(\log(n)/n^{1-1/q}) \quad (27)$$

and

$$P_{ON}(n) = O(\log(n)/n). \quad (28)$$

It is known [11] that even in the case of finite decision space \mathcal{X} , and known parametric noise having moment generating functions, the best achievable rate for on-line strategies is $P[\text{Mistake}(n)] = O(1/n)$.

III. CONCLUSIONS

The present note contributes to the important decision-theoretic issue of how to undertake experimental design during an exploration phase so that the terminal decision is as wise as possible. This is an important piece of the larger picture of machine learning, a rigorous "black-box" approach to asymptotically optimal decision making.

From applications, e.g., [12]–[14], there is a convincing case that our machine learning techniques are feasible and useful for problems for which alternatives (such as genetic algorithms and neural nets) do not yet seem viable. Much of the theory is in place, and in particular, there are results about what is achievable [10], [11], [15], [16] and algorithms for attainment of the achievable. Some nontrivial toy problems have been mastered and initial investigations on serious problems (AIDS policy [13] and assembly line control [17]) have been published. We foresee a rich future for this statistical methodology. Ongoing investigations include algorithms for identification, control, and prediction for nonlinear systems to complement methodology of linear systems.

The present note, which is our first foray into off-line strategies, has contributed to the understanding of convergence behavior. We have noted that Gurin, Devroye, and others had already investigated off-line algorithms and established conditions for convergence. But to our knowledge, we are the first to "tune" these strategies to consider rates and trade-off between allocation of times to acquiring and testing new samples *vs* refining estimates of samples already collected. Clearly, algorithms in this note are to be preferred to the on-line methods when there is an experimental or simulation phase of system tuning during which actual performance is of no concern. The methods offered in Section II are optimal, in a minimax rate sense, with respect to asymptotic convergence, but they are practical, and easy to code. The memory requirements are small: at time n , one needs only to have stored the test points, the sample means, and the sample counts, i.e., the triplets $\{t(i), m_i, NS(i)\}$: $1 \leq i \leq NP(i)$. Even for a huge number of algorithm iterations, and a fairly large decision-space dimension, this is a modest demand because $NP(n)$ grows at most as \sqrt{n} . In many cases, the time to the final mistake can be seen [18] to have a short tail. In brief, the claim is that these learning algorithms are simple, efficient, and plainly feasible.

ACKNOWLEDGMENT

This note was strongly influenced by conversations, publications, and encouragement of L. Devroye, and fills out an invited (by Devroye) presentation [19] at the 1992 Winter Simulation Conference

REFERENCES

- [1] S. Yakowitz and M. Kollier, Machine learning with application to counting strategies for blackjack, *J. Statist. Plann. and Inference*, vol. 13, pp. 295–309, 1992.
- [2] L. P. Devroye, On random search with a learning memory, in *Proc. 9th IEEE Conf. Cybernetics Society*, pp. 704–711, 1976.
- [3] ———, The uniform convergence of nearest neighbor regression function estimators and their application to optimization, *IEEE Trans. Inform. Theory*, vol. 24, pp. 142–151, 1978.
- [4] R. A. Jarvis, Optimization strategies in adaptive control: A selective survey, *IEEE Trans. Syst. Man Cybernet.*, vol. SMC-5, pp. 83–94, 1975.
- [5] I. Matyas, Random optimization, *Automat. Remote Contr.*, vol. 26, pp. 244–251, 1965.
- [6] I. S. Gurn, Random search in the presence of noise, *Engin. Cybernet.*, vol. 4, no. 3, pp. 252–260, 1966.
- [7] I. Wagner, On the rate of convergence for the law of large numbers, *Annals Math. Statist.*, vol. 40, pp. 2195–2197, 1969.
- [8] J. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*. New York: Wiley, 1990.
- [9] M. Csörgő, *Quantile Processes with Statistical Applications*. Philadelphia, PA: SIAM, 1983.
- [10] S. Yakowitz and W. Lowe, Nonparametric bandits, *Ann. Operat. Res.*, vol. 28, pp. 297–312, 1991.
- [11] T. L. Lai and H. Robbins, Asymptotically efficient adaptive allocation rules, *Adv. Appl. Math.*, vol. 6, pp. 4–22, 1985.
- [12] S. Yakowitz, A statistical foundation for machine learning with application to gomoku, *Comp. Math. Appl.*, vol. 17, pp. 1095–1107, 1989.
- [13] ———, A decision model and methodology for the AIDS epidemic, *Appl. Math. Computat.*, vol. 52, pp. 149–172, 1992.
- [14] S. Yakowitz, R. Hayes, and J. Gani, Automatic learning for dynamic Markov random fields with application to epidemiology, *Operations Res.*, vol. 40, pp. 867–876, 1992.
- [15] S. Yakowitz and E. Lugosi, Random search in the presence of noise with application to machine learning, *SIAM J. Statist. Comput.*, vol. 11, no. 4, pp. 702–712, 1990.
- [16] S. Yakowitz, A globally convergent stochastic approximation, *SIAM J. Contr. Optim.*, vol. 31, pp. 30–40, 1993.
- [17] S. Yakowitz, T. Jayawardena, and S. Li, Theory for automatic learning under partially observed Markov dependent noise, *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1316–1324, 1992.
- [18] J. Pinelis and S. Yakowitz, The time until the final zero crossing of random sums with application to nonparametric bandit theory, *Appl. Math. Computat.*, vol. 63, pp. 235–256, 1994.
- [19] S. Yakowitz, Automatic learning: theorems for concurrent simulation and optimization, in *Proc. 1992 Winter Simulation Conf.*, J. J. Swain, D. Goldsman, R. C. Crun, and J. R. Wilson, eds., Arlington, VA, 1992, pp. 487–493.

Discrete-Time Observers with Random Noises in Dynamic Block

by A. Lyashenko and L. B. Ryashko

Abstract—The minimum variance state estimation of linear stochastic discrete-time systems by an observer of reduced-order is investigated. There is an additional random noise with known intensity in the dynamic block of the observer. The local optimal reduced-order state estimator is found which takes into account the presence of such noise. The equations of an optimal stationary observer are derived for linear stochastic time invariant system.

1. INTRODUCTION

We consider the following discrete time linear system

$$x_{t+1} = Ax_t + \xi_t \quad (1)$$

$$y_t = Cx_t + \eta_t \quad (2)$$

where $x_t \in R^n$, $y_t \in R^m$ are the state and the measurement vectors, respectively. A and C are matrices of appropriate dimensions ($\text{rank}(C) = m$), the vectors ξ_t and η_t are zero mean discrete time noise signals that satisfy

$$E\xi_t \xi_t^T = W > 0, \quad E\eta_t \eta_t^T = V > 0 \quad (3)$$

We assume also that x_0 is Gaussian with mean and covariance given by

$$Ex_0 = \bar{x}_0, \quad E(x_0 - \bar{x}_0)(x_0 - \bar{x}_0)^T = \Delta_0 \quad (4)$$

Furthermore x_0 , ξ_t and η_t are mutually independent.

An optimal (minimum variance) estimate \hat{x}_t of x_t and its covariance Δ_t are usually formed by the discrete Kalman-Bucy filter (KBF). In practice, the presence of disregarded errors connected with inaccuracy in the model used or with roundoff errors in the computer during digital computations can easily lead to entirely unacceptable estimates and moreover to divergence of KBF [1]–[3]. In [10] such errors were modeled by random disturbances in the filter. An optimal estimator was constructed which takes into account the influence of this additional noise. In some cases, it enables us to improve essentially the quality of the estimates obtained.

The present note that deals with similar investigation of reduced order observer (ROO) with additional noise in the dynamic block goes along with vast research into robust estimation [4]–[5]. It turns out that the estimates of ROO may be much better than the full order filter's ones.

This note is structured as follows. In Section II we design observers for three different situations: ROO 1—there is no additional noise; ROO 2—the presence of noise in the dynamic block is ignored; ROO 3—the influence of such noise is taken into account by seeking out the parameters of an optimal observer. The error equations are given for all of these estimators. In Section III we consider the design of an optimal stationary observer with additional noise for time invariant systems. The equations for its parameters are presented. Section IV contains an example that demonstrates the stability of ROO in the presence of noise in the dynamic block.

Manuscript received September 16, 1993; revised March 27, 1994.

The authors are with Ural State University, 620083 Ekaterinburg, Russia. IEEE Log Number 9406980.

II. REDUCED-ORDER OBSERVERS WITH ADDITIONAL NOISES AND THEIR ERRORS

Define a reduced-order observer for system (1) as a pair of dynamic block

$$\dot{z}_{i+1} = F_i z_i + D_i y_i + \nu_{i+1} \quad (5)$$

and connection

$$\dot{\hat{x}}_i = P_i \dot{z}_i + R_i y_i. \quad (6)$$

Here $z_i \in R^l$, $\hat{x}_i \in R^n$ are the estimation vectors for $T_i x_i$ and x_i , respectively; T_i , F_i , D_i , P_i , R_i are matrices of appropriate dimensions; $\nu_i \in R^n$ is a zero mean discrete-time random signal with covariance $E\nu_i \nu_i^T = Q_i$, that is independent of other system noises.

Case $\nu_i = 0$ has been completely investigated in [6]. We assume as usual that parameters of ROO satisfy the Huddle structure

$$F_i = T_{i+1} A_i P_i, \quad D_i = T_{i+1} A_i R_i \quad (7)$$

and the limitation on precision

$$P_i T_i + R_i C_i = I_n. \quad (8)$$

With respect to (7), (8), the error matrix $\Delta_i = E(x_i - \hat{x}_i)(x_i - \hat{x}_i)^T$ of estimate \hat{x}_i is given by the recursive relation

$$\Delta_{i+1} = P_{i+1} T_{i+1} \Pi_i(\Delta_i) T_{i+1}^T P_{i+1}^T + R_{i+1} V_{i+1} R_{i+1}^T + P_{i+1} Q_{i+1} P_{i+1}^T \quad (9)$$

where

$$\Pi_i(\Delta) = A_i \Delta A_i^T + W_{i+1}.$$

Let the first m_1 components of measurement vector (2) be disturbed by comparatively small noises. These measurements would be used for reducing the order of an observer to $r = n - m_1$. We assume also, without loss of generality, that system (1), (2) has the observation canonic form

$$\begin{bmatrix} x_{i+1}^1 \\ x_{i+1}^2 \end{bmatrix} = \begin{bmatrix} A_i^{11} & A_i^{12} \\ A_i^{21} & A_i^{22} \end{bmatrix} \begin{bmatrix} x_i^1 \\ x_i^2 \end{bmatrix} + \begin{bmatrix} \xi_{i+1}^1 \\ \xi_{i+1}^2 \end{bmatrix},$$

$$y_i^1 = x_i^1 + \eta_i^1, \quad y_i^2 = \dot{C} x_i^2 + \eta_i^2. \quad (10)$$

Here, x_i^1 , y_i^1 , ξ_i^1 , η_i^1 are m_1 -dimensional vectors; x_i^2 , ξ_i^2 are $(n - m_1)$ -dimensional vectors and y_i^2 , η_i^2 are $(m - m_1)$ vectors. The matrices A_i , C_i , and V_i are partitioned according to

$$A_i = \begin{bmatrix} A_i^{11} & A_i^{12} \\ A_i^{21} & A_i^{22} \end{bmatrix}, \quad C_i = C' = \begin{bmatrix} I_{m_1} & 0 \\ 0 & \dot{C}' \end{bmatrix}, \quad V_i = \begin{bmatrix} V_i^{11} & V_i^{12} \\ V_i^{21} & V_i^{22} \end{bmatrix}$$

respectively, where A_i^{11} , V_i^{11} are $m_1 \times m_1$; A_i^{12} is $m_1 \times (n - m_1)$; A_i^{21} is $(n - m_1) \times m_1$; A_i^{22} is $(n - m_1) \times (n - m_1)$; $\dot{C}' = [I_{m-m_1} : 0]$ is $(m - m_1) \times (n - m_1)$; $V_i^{12} = V_i^{21T}$ is $m_1 \times (m - m_1)$; V_i^{22} is $(m - m_1) \times (m - m_1)$.

Since η_i^1 is small, we adopt $x_i^1 = y_i^1$ that entails (see (6)–(8)) the following:

$$P_i = \begin{bmatrix} 0 \\ I_{n-m_1} \end{bmatrix}, \quad R_i = \begin{bmatrix} I_{m_1} & 0 \\ R_i^1 & R_i^2 \end{bmatrix}, \quad (11)$$

$$T_i = [-R_i^1 : I_{n-m_1} - R_i^2 \dot{C}'],$$

where R_i^1 is $(n - m_1) \times m_1$ and R_i^2 is $(n - m_1) \times (m - m_1)$. So, if we know Δ_i , then the error matrix Δ_{i+1} is entirely defined by $\hat{R}_{i+1} = [R_{i+1}^1 : R_{i+1}^2]$ and satisfies the equation

$$\Delta_{i+1} = \Phi_{i+1}(\Delta_i, \hat{R}_{i+1}) + S_{i+1}$$

where

$$\Phi_{i+1}(\Delta, \hat{R}) = (I_n - RC)\Pi_i(\Delta)(I_n - RC)^T + R V_{i+1} R^T,$$

$$R = \begin{bmatrix} I_{m_1} & 0 \\ R^1 & R^2 \end{bmatrix}, \quad S_{i+1} = \begin{bmatrix} 0 & 0 \\ 0 & Q_{i+1} \end{bmatrix}. \quad (12)$$

An optimal estimator is derived from the following problem:

$$J = \text{tr} \Phi_{i+1}(\Delta_i, \hat{R}_{i+1}) \rightarrow \min_{\hat{R}_{i+1}}.$$

Let there be no noise in the dynamic block of ROO ($\nu_i = 0$). Then the parameters of an optimal observer are defined as follows:

$$\hat{R}_{i+1}^{(1)} = \Psi_{i+1}(\Delta_i^{(1)}). \quad (13)$$

$$\Delta_{i+1}^{(1)} = \Phi_{i+1}(\Delta_i^{(1)}), \quad \Delta_0^{(1)} = \Delta_0 \quad (14)$$

where

$$\Psi_{i+1}(\Delta) = [\Pi_i^{21}(\Delta) : \Pi_i^{22}(\Delta) \dot{C}^T] (C \Pi_i(\Delta) C^T + V_{i+1})^+. \quad (15)$$

$$\Phi_i = \begin{bmatrix} \Phi_i^{11} & \Phi_i^{12} \\ \Phi_i^{21} & \Phi_i^{22} \end{bmatrix},$$

$$\Phi_i^{11} = V_i^{11}, \quad \Phi_i^{12} = [V_i^{11} : V_i^{12}] \Psi_i^T(\Delta), \quad (16)$$

$$\Phi_i^{21} = \Phi_i^{12T},$$

$$\Phi_i^{22} = \Pi_{i-1}^{22}(\Delta) - [\Pi_{i-1}^{21}(\Delta) : \Pi_{i-1}^{22}(\Delta) \dot{C}^T] \Psi_i^T(\Delta).$$

Similar equations were derived in [6]. The observer which satisfies (13)–(16), will be denoted ROO-1. The estimate $\hat{x}_i^{(1)}$, of ROO-1 has the error $\Delta_i^{(1)}$.

Now, we consider the estimator (5), (6) with disturbances $\nu_i \neq 0$. Let the matrix \hat{R}_{i+1} be given as before by (13)–(16): $\hat{R}_{i+1}^{(2)} = \hat{R}_{i+1}^{(1)}$. Such construction will be called ROO-2 and its estimate will be designated as $\hat{x}_i^{(2)}$. The matrix of real error $\Delta_i^{(2)}$ is connected with the expected error $\Delta_i^{(1)}$ according to

$$\Delta_i^{(2)} = \Delta_i^{(1)} + G_i. \quad (17)$$

The additive matrix G_i , caused by noise ν_i , has the following structure:

$$G_i = \begin{bmatrix} 0 & 0 \\ 0 & \dot{G}_i \end{bmatrix}$$

where $(n - m_1) \times (n - m_1)$ matrix \dot{G}_i is calculated from the recurrent equation

$$\dot{G}_{i+1} = B_{i+1} \dot{G}_i B_{i+1}^T + Q_{i+1}, \quad \dot{G}_0 = 0, \quad (18)$$

$$B_{i+1} = (I_n - R_{i+1}^{(1)} C') A_i.$$

Obviously, ROO-2 is not optimal because we ignore the information about the additional noise when we choose the coefficient \hat{R}_{i+1} .

ROO 3 that optimally uses such information for the selection R_{+1} satisfies

$$R_{+1}^{(i)} = \Psi_{+1}(\Delta_{+1}^{(i)}) \quad (19)$$

$$\Delta_{+1}^{(i)} = \Phi_{+1}(\Delta_{+1}^{(i-1)}) + S_{+1} \quad \Delta_0^{(i)} = \Delta_0 \quad (20)$$

The next lemma proves the optimum of ROO 1 and ROO 3

Lemma Let R be an arbitrary $n \times m$ matrix that satisfies (11). Then

$$\Delta_{+1}(R) = \Delta_{+1}^{(i)} + (R - R_{+1}^{(i)})[C\Pi(\Delta)(I - RC)^T + V_{+1}] \\ (R - R_{+1}^{(i)})^T + I_{+1}(R - R_{+1}^{(i)}) \quad (21)$$

where

$$I_{+1} = \begin{bmatrix} 0 & I_{+1}^T \\ I_{+1} & 0 \end{bmatrix}$$

$$I_{+1}^T = [V_{+1}^T \quad V_{+1}^T](R - R_{+1}^{(i)})^T \quad I_{+1} = I_{+1}^T$$

The proof follows directly from the relations (12)–(15)–(16)–(19) and (20). This lemma is correct for ROO 1 when $\nu = 0$.

Note ROO 1–3 are only local optimal observers. Actually the values of R_{+1} are chosen step by step independently to minimize $J = \text{tr} \Delta_{+1}$ with given Δ . It may be easily shown that the joint selection of matrices R_1, R_2, \dots, R_{+1} enables us to obtain a smaller value of $\text{tr} V_{+1}$. We remark that both procedures yield the same result for an observer of full order (in this case $m_1 = 0, I = 0$).

III. STATIONARY REDUCED ORDER OBSERVERS

We consider the following time invariant system

$$\dot{x}_{+1} = Ax_{+1} + \xi_{+1} \quad (22)$$

$$y = Cx_{+1} + \eta$$

where

$$C = \begin{bmatrix} I & 0 \\ 0 & C \end{bmatrix} \quad C = [I \quad 0]$$

$$E\xi\xi^T = W > 0 \quad E\eta\eta^T = V \geq 0$$

and the stationary observer

$$\dot{z} = I\dot{z} + Dy + z_{+1} \quad (23)$$

$$z = P\dot{z} + Ry$$

where

$$P = \begin{bmatrix} 0 \\ I \end{bmatrix} \quad R = \begin{bmatrix} I & 0 \\ R^T & R \end{bmatrix} \quad R = [R^T \quad R]$$

$$\Gamma = TAP - D - TARB - T = [-R^T \quad I \quad 0 \quad -RC]$$

The constant matrix R is chosen from the set $\mathbb{R} = \{R/(I - RC)A - \text{stable}\}$. For any $R \in \mathbb{R}$ the sequence of matrices $\Delta(R)$ from (12) has the limit $\Delta(R) = \lim_{i \rightarrow \infty} \Delta_{+1}^{(i)}(R)$ which satisfies the Lyapunov matrix equation

$$\Delta = (I - RC)\Pi(\Delta)(I - RC)^T + R[V \quad R^T] + S \\ \Pi(\Delta) = A\Delta A^T + W \quad (24)$$

In such circumstances the problem of the design of an optimal stationary ROO may be formulated as follows

$$\text{tr} \Delta \rightarrow \min_{R \in \mathbb{R}} \quad (25)$$

where Δ satisfies (24). Lagrange function for this problem

$$H = H(\Delta, \Lambda, R) = \text{tr} \Delta \\ + \text{tr}[(\Delta - (I - RC)\Pi(\Delta)(I - RC)^T + V)]$$

where

$$\Delta = \begin{bmatrix} \Delta^{11} & \Delta^T \\ \Delta^T & \Delta \end{bmatrix} \quad \Lambda = \begin{bmatrix} \Lambda^{11} & \Lambda^T \\ \Lambda^T & \Lambda \end{bmatrix}$$

are the error matrix of stationary observer and the matrix of Lagrange multipliers, respectively

$$\Pi(\Delta) = \begin{bmatrix} \Pi^{11} & \Pi^T \\ \Pi^T & \Pi \end{bmatrix}$$

Here $\Delta^{11}, \Lambda^{11}, \Pi^{11}$ are $m_1 \times m_1$, $\Delta^T = \Delta^{1T} = \Lambda^{1T} = \Lambda^{1T}$, $\Pi^T = \Pi^{1T}$ are $m_1 \times (n - m_1)$, $\Delta = \Lambda = \Pi$ are $(n - m_1) \times (n - m_1)$. The necessary conditions of extremum are presented by the nonlinear system

$$\Delta^{11} - \Lambda^{11} = 0$$

$$\Delta^T - \Lambda^{1T}R^{1T} - \Lambda^T R^T = 0 \quad (26)$$

$$\Delta = R^T(\Pi^T + V^{11})R^{1T} - R\Lambda^{1T}R^T - R\Lambda - R \\ - R\Lambda - R^T + (I - RC) - R(C)\Pi - R \\ + R^T\Pi^T(I - RC)^T - R(C)^T - Q = 0$$

$$I + \Lambda^{11} - [R^T A^{11} - (I - RC)A^{1T}]^T \\ \Lambda = [R^T A^{11} - (I - RC)A^{1T}] = 0$$

$$\Lambda^{1T} - [R^T A^{11} - (I - RC)A^{1T}]^T \\ \Lambda = [R^T A^{11} - (I - RC)A^{1T}] = 0 \quad (27)$$

$$I - \Lambda^{11} + \Lambda = [R^T A^{11} - (I - RC)A^{1T}]^T \\ \Lambda = [R^T A^{11} - (I - RC)A^{1T}] = 0$$

$$\Lambda = [R^T(\Pi^{11} + V^{11}) + R(C\Pi^T + V^T) - \Pi] \\ + \Lambda^{11}\Lambda^{11} = 0 \quad (28)$$

$$\Lambda = [R^T(\Pi^T C^T + V^T) + R(C\Pi C^T + V) - \Pi C^T] \\ + \Lambda^{11}\Lambda^{11} = 0$$

The equations (26)–(28) give us the stationary observers that are analogous to ROO 1–3 from Section II. The error of stationary ROO 2 is defined from (24) by substitution $R^{(i)} = R^{(1)}$. Section IV contains the comparison of stationary observers 1–2–3.

Note For system (22) as in [9] we may find the limiting ($i \rightarrow \infty$) values of the observers from Section II. These limiting estimators will differ from the corresponding stationary ones of Section III. We remark that such difference does not take place when $\eta^1 = 0$ (first m_1 measurements without noises) and in case of full order [10].

IV. EXAMPLE

We consider the following differential system:

$$\dot{x} = \hat{A}x(t) + b\dot{\xi}(t) \quad (29)$$

where

$$\hat{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad x(t) = \begin{bmatrix} x^1(t) \\ x^2(t) \end{bmatrix}$$

$$E\dot{\xi}(t) = 0, \quad E\dot{\xi}(t)\dot{\xi}(\tau) = \delta(t - \tau)$$

with discrete measurements in steady time moments $t_{i+1} = t_i + h$

$$y_i = [1 \ 0]x_i. \quad (30)$$

The state vector $x_i = x(t_i)$ of continuous-time system (29) satisfies the following discrete system:

$$x_{i+1} = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix} x_i + \xi_{i+1} \quad (31)$$

where

$$\xi_{i+1} = \int_{t_i}^{t_{i+1}} \exp[\hat{A}(t_{i+1} - \tau)] b d\dot{\xi}(\tau)$$

$$E\xi_i = 0, \quad E\xi_i \xi_i^T = \begin{bmatrix} h^3/3 & h^2/2 \\ h^2/2 & h \end{bmatrix}.$$

We compare the estimation precision of Kalman filter [10]

$$\hat{x}_{i+1} = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix} \hat{x}_i + \begin{bmatrix} 1 \\ k \end{bmatrix} (y_{i+1} - [1 \ h] \hat{x}_i) + \begin{bmatrix} \nu_{i+1}^1 \\ \nu_{i+1}^2 \end{bmatrix} \quad (32)$$

and the reduced-order observer (23), (24)

$$z_{i+1} = (1 - hr)z_i - y_i hr^2 + \nu_{i+1}, \quad (33)$$

$$\hat{x}_{i+1} = \begin{bmatrix} y_{i+1} \\ z_{i+1} + r y_{i+1} \end{bmatrix}$$

which contain the mutually independent noises ν_i^1 , ν_i^2 , and ν_i with characteristics $E\nu_i^1 = E\nu_i^2 = E\nu_i = 0$, $E(\nu_i^1)^2 = E(\nu_i^2)^2 = E(\nu_i)^2 = q$. We shall take the magnitude $\delta = \lim_{h \rightarrow 0} E(x_i^2 - \hat{x}_i^2)^2$ as a precision test and denote the error of KBF as δ_f , and the error of ROO as δ_o .

1): Let $\nu_i^1 = \nu_i^2 = \nu_i = 0$ (there is no additional noise). Then the optimal parameters of KBF-1 and ROO-1 are equivalent

$$k = r = \frac{3 + \sqrt{3}}{h(2 + \sqrt{3})}, \quad \delta_f^{(1)} = \delta_o^{(1)} = \frac{h}{2\sqrt{3}}. \quad (34)$$

2): Let the filter and the observer be disturbed by noise but the parameters r and k be chosen as in (34) (the presence of noise is ignored). Then

$$\delta_o^{(2)} = \delta_o^{(1)} + \frac{q}{4\sqrt{3} - 6}, \quad \delta_f^{(2)} = \delta_f^{(1)} + \frac{q\sqrt{3}}{h^2}. \quad (35)$$

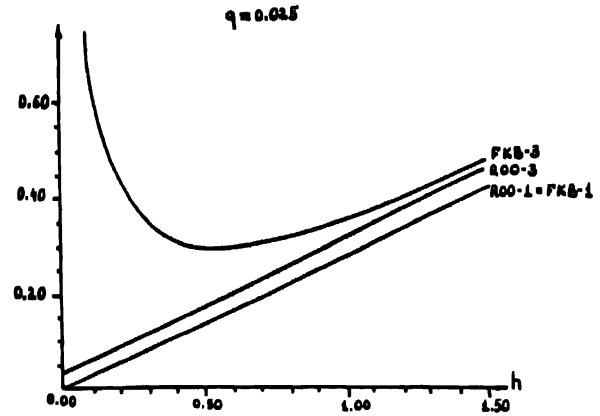


Fig. 1. The errors of ROO-1, 3 and KBF-1, 3 (ROO-1 = KBF-1).

3): The parameters of KBF and ROO are selected with account of additional noise. So, we obtain

$$r = \frac{\delta_o^{(3)} + h/2}{h(\delta_o^{(3)} + h/3)}, \quad \delta_o^{(3)} = \frac{1}{2} \left(q + \sqrt{q^2 + \frac{h}{3}(4q + h)} \right),$$

$$k = \frac{\delta_f^{(1)} + h/2}{h(\delta_f^{(1)} + h/3) + q},$$

$$\delta_f^{(1)} = \frac{1}{2} \left(q + \sqrt{q^2 + \frac{h}{3}(4q + h) + \frac{4q}{h} \left(\frac{q}{h} + 1 \right)} \right). \quad (36)$$

The error behavior of the optimal stationary filter and observer with noises depending on the discretization step h of system (29) is shown in Fig. 1.

If there is no additional noise then the estimation errors of KBF and ROO are the same and are improved by decrease of step h : $\delta_f^{(1)} = \delta_o^{(1)} = O(h)$. The presence of noise makes the estimation error of KBF essentially larger when h is small: $\delta_f^{(2)} = O(1/h^2)$. The optimization of KBF only lowers the order of vanishing: $\delta_f^{(1)} = O(1/h)$. The errors of ROO-2, 3 keep in the natural monotony on h .

We can correct KBF-2, 3 by the suitable selection of step h [7]. However, the Fig. 1 demonstrates that this method is less effective than the use of corresponding ROO. It may be also shown that the employment of a reduced-order observer is more preferable here than the known regularization [8] of the Kalman filter.

V. CONCLUSION

In this note, we consider the state estimation of linear discrete-time stochastic system by a reduced-order observer with noise in the dynamic block. It turns out that the estimation quality of ROO may be much better than the quality of full-order filter. We derived the local optimal observer which takes into account the influence of additional noise. The optimal stationary observer was designed for time-invariant systems. An example of Section IV demonstrates the stability of ROO in the presence of noise in the dynamic block.

REFERENCES

- [1] R. J. Fitzgerald, "Divergence of Kalman filter," *IEEE Trans. Automat. Contr.*, vol. AC-16, no. 6, pp. 736-747, 1971.
- [2] D. Williamson, "Finite wordlength design of digital Kalman filters for state estimation," *IEEE Trans. Automat. Contr.*, vol. AC-30, no. 10, pp. 930-939, 1985.
- [3] A. P. Sage and J. L. Melsa, *Estimation Theory with Application to Communications and Control*. New York: McGraw-Hill, 1971.
- [4] P. Dorato, "A historical review of robust control," *IEEE Contr. Syst. Mag.*, vol. 7, pp. 44-47, 1987.

- [5] E. Yaz, "Robustness of stochastic parameter control and estimation schemes," *IEEE Trans Automat Contr*, vol. AC-35, no. 5, pp. 637-640, 1990.
- [6] C. T. Leondes, ed., *Control and Dynamic Systems*. New York: Academic, 1976.
- [7] G. N. Milshstein and S. A. Pjanzin, "Digital modeling of a Kalman-Bucy filter and an optimal filter in quantized arrival of information," *Automat Remote Contr*, no. 1, pp. 59-68, 1985.
- [8] R. Sh. Liptser, "Equations of near optimal Kalman filter with a singular matrix of noise covariations in observations," *Automat Remote Contr*, no. 1, pp. 35-41, 1974.
- [9] A. V. Balakrishnan, *Kalman Filtering Theory*. New York: Optimization Software Inc. Publication Division, 1984.
- [10] E. A. Lyashenko and I. B. Ryashko, "On the estimation by means of filter containing random noises," *Automat Remote Contr*, no. 2, pp. 75-83, 1992.

Nevanlinna-Pick Interpolation Problem for Two Frequency Scale Systems

Hossein M. Oloomi

Abstract—The Nevanlinna-Pick interpolation problem for a parametric set of data is considered. Sufficient conditions are obtained that guarantee a two frequency scale solution. Decomposition of the problem into two Nevanlinna-Pick interpolation problems with smaller data size is presented. An algorithm is developed which computes the solution as the result of combining the solutions for these two smaller problems.

I. INTRODUCTION

Many interesting problems of circuits and systems, as well as a number of control problems, can be formulated as the Nevanlinna-Pick (NP) interpolation problem [1]. Indeed, many early attempts to solve the H_∞ problem used this interpolation technique as its mathematical foundation. Examples include [2] and [3]. The NP interpolation problem was also used to solve the robust stabilization problem in [4] and [5]. A good account of the theory is given in [6]–[8]. We use the version presented in [4].

The NP problem considered in this note is different from the standard NP problem in a number of ways. The main difference is in the data set. The data considered here is more general because it is parametric. This should be contrasted with the standard case where data is a fixed set of numbers. However, only first order singularity and analytic perturbation are allowed. This particular structure was considered since it enables us to seek a two frequency scale solution. Nonetheless, generalization to higher order singularities is possible provided one allows multifrequency scale solutions.

Two frequency scale (TFS) transfer functions were first introduced in [9]. The most notable feature of a TFS transfer function is that it can be studied in terms of two lower order transfer functions referred to as the slow transfer function and the fast transfer function. It is evident that when complexity and high dimensionality is of major concern, TFS property is a highly useful attribute. This is precisely the motivation behind the present note.

Once it is guaranteed that the NP problem yields a TFS solution, it certainly makes sense to approximate the solution by constructing

another nearby solution which is easier to obtain than the solution which is obtained as the result of solving two smaller NP problems and is close to the actual solution in some norm. This idea will be pursued later in this note.

We refer the interested reader to [9] for the definition of TFS transfer functions. We will denote the set of all TFS transfer functions by T . When $u(s, \epsilon) \in T$, we will denote its slow and fast transfer functions by $u^s(s)$ and $u^f(p)$ respectively, where $s = j\omega$. We will also need the definition of strictly bounded real functions which we will give here for completeness.

Definition 1 A function $u(s)$ analytic in $\text{Re}[s] > 0$ is said to be a strictly bounded real (SBR) function if $|u(j\omega)| < 1$ for all ω .

The following lemma will be needed in the sequel.

Lemma 1 Let $u(s, \epsilon) \in T$ have no unstable lost poles and have SBR slow and fast transfer functions $u^s(s)$ and $u^f(p)$ respectively. Then for sufficiently small ϵ , $u(s, \epsilon)$ is an SBR transfer function.

Q.E.D.

In [10], it has been proven that if $u(s, \epsilon) \in T$ has no unstable lost poles and if $\|u^s(s)\|_\infty \leq 1$ and $\|u^f(p)\|_\infty \leq 1$ then for sufficiently small ϵ , $\|u(s, \epsilon)\|_\infty \leq 1 + O(\epsilon)$. We specialize this result to the present situation. We have $\|u^s(s)\|_\infty = \theta^s < 1$ and $\|u^f(p)\|_\infty = \theta^f < 1$. Let $\theta_{\text{eff}} = \max\{\theta^s, \theta^f\}$. Then there exists a constant K such that for sufficiently small ϵ , $\|u(s, \epsilon)\|_\infty \leq \theta_{\text{eff}} + K\epsilon$. When $K \leq 0$, then $\|u(s, \epsilon)\|_\infty \leq \theta_{\text{eff}} + K\epsilon < \theta_{\text{eff}} < 1$. On the other hand, when $K > 0$, set $\epsilon^* = \frac{1 - \theta_{\text{eff}}}{K} > 0$. Then for all $0 \leq \epsilon < \epsilon^*$, $\|u(s, \epsilon)\|_\infty \leq \theta_{\text{eff}} + K\epsilon < \theta_{\text{eff}} + K\left(\frac{1 - \theta_{\text{eff}}}{K}\right) = 1$. Hence $\|u(s, \epsilon)\|_\infty < 1$, i.e., for sufficiently small ϵ , $u(s, \epsilon)$ is an SBR function.

Q.E.D.

II. THE INTERPOLATION PROBLEM

Suppose we are given two sequences of l complex numbers

$$\begin{aligned} \alpha_1(\epsilon) &= \alpha_{l_1}(\epsilon) = \frac{\alpha_{l_1+1}(\epsilon)}{\epsilon} = \dots = \alpha_l(\epsilon) \\ \beta_1(\epsilon) &= \beta_{l_1}(\epsilon) = \beta_{l_1+1}(\epsilon) = \dots = \beta_l(\epsilon) \end{aligned}$$

where each $\alpha_i(\epsilon)$ and $\beta_i(\epsilon)$ is a function of ϵ and analytic at $\epsilon = 0$. In addition, suppose there exists an $\epsilon^* > 0$ such that for all $0 \leq \epsilon < \epsilon^*$, $\text{Re}[\alpha_i(\epsilon)] > 0$ and $|\beta_i(\epsilon)| < 1$ for $i = 1, \dots, l$. The NP problem for this set of data is one of finding an SBR function $u(s, \epsilon)$ which interpolates the data, i.e.,

$$u(\alpha_i(\epsilon), \epsilon) = \beta_i(\epsilon) \quad i = 1, \dots, l$$

It is well known that the solution exists (see [4]) iff the Pick matrix

$$P(\epsilon) = \begin{bmatrix} P_{11}(\epsilon) & \epsilon P_{1l}(\epsilon) \\ \epsilon \overline{P_{1l}(\epsilon)} & \epsilon P_{ll}(\epsilon) \end{bmatrix} > 0$$

where

$$\begin{aligned} P_{11}(\epsilon) &= \left\{ \frac{1 - \beta_j(\epsilon) \overline{\beta_i(\epsilon)}}{\alpha_i(\epsilon) + \overline{\alpha_j(\epsilon)}} \right\} & i = 1 & \dots & l_1 \\ & & j = 1 & \dots & l_1 \\ P_{ll}(\epsilon) &= \left\{ \frac{1 - \beta_j(\epsilon) \overline{\beta_i(\epsilon)}}{\epsilon \alpha_i(\epsilon) + \overline{\epsilon \alpha_j(\epsilon)}} \right\} & i = l_1 + 1 & \dots & l \\ & & j = l_1 + 1 & \dots & l \\ P_{22}(\epsilon) &= \left\{ \frac{1 - \beta_j(\epsilon) \overline{\beta_i(\epsilon)}}{\alpha_i(\epsilon) + \overline{\alpha_j(\epsilon)}} \right\} & i = l_1 + 1 & \dots & l \\ & & j = l_1 + 1 & \dots & l \end{aligned}$$

Manuscript received October 4, 1993; revised June 7, 1994.
The author is with the Department of Electrical Engineering, Purdue University at Fort Wayne, Fort Wayne, IN 46805 USA.
IEEE Log Number 9406981.

Moreover, the solution $u(s, \epsilon) := u_1(s, \epsilon)$ can be generated recursively from the algorithm (see [7])

$$u_j(s, \epsilon) := \frac{(\epsilon s - \alpha_j(\epsilon))u_{j+1}(s, \epsilon) + \rho_j(\epsilon)(\epsilon s + \bar{\alpha}_j(\epsilon))}{(\epsilon s + \bar{\alpha}_j(\epsilon)) + \bar{\rho}_j(\epsilon)(\epsilon s - \alpha_j(\epsilon))u_{j+1}(s, \epsilon)},$$

$$j = l, \dots, l_1 + 1, \quad (1)$$

$$u_j(s, \epsilon) := \frac{(s - \alpha_j(\epsilon))u_{j+1}(s, \epsilon) + \rho_j(\epsilon)(s + \bar{\alpha}_j(\epsilon))}{(s + \bar{\alpha}_j(\epsilon)) + \bar{\rho}_j(\epsilon)(s - \alpha_j(\epsilon))u_{j+1}(s, \epsilon)},$$

$$j = l_1, \dots, 1 \quad (2)$$

where $u_{l+1}(s, \epsilon)$ is any SBR function, and the Fenyves array $\beta_{i,j}(\epsilon)$ is generated via

$$\beta_{i,1}(\epsilon) := \beta_i(\epsilon), \quad i = 1, \dots, l,$$

$$\beta_{i,j+1}(\epsilon) := \frac{(\alpha_i(\epsilon) + \bar{\alpha}_j(\epsilon))(\beta_{i,j}(\epsilon) - \beta_{j,j}(\epsilon))}{(\alpha_i(\epsilon) - \alpha_j(\epsilon))(1 - \bar{\beta}_{j,j}(\epsilon)\beta_{i,j}(\epsilon))},$$

$$1 \leq j \leq i-1 \leq l-1,$$

$$\rho_j(\epsilon) := \beta_{j,j}(\epsilon), \quad j = 1, \dots, l.$$

It is well-known that $P(\epsilon) > 0$ iff $|\beta_{i,j}(\epsilon)| < 1$, $\forall i, j$ [7]. In particular, $P(\epsilon) > 0$ implies $|\rho_j(\epsilon)| < 1$, $\forall j$. We call this the full-order NP problem. Associated with this problem, we consider two subproblems.

1) The Fast NP Problem: Consider two sequences of $(l-1)$ complex numbers

$$\alpha_{l_1+1}^f, \dots, \alpha_l^f$$

$$\beta_{l_1+1}^f, \dots, \beta_l^f$$

with the property that $\operatorname{Re}[\alpha_i^f] > 0$ and $|\beta_i^f| < 1$ for $i = l_1+1, \dots, l$. We seek to find an SBR function $u^f(p)$ such that

$$u^f(\alpha_i^f) = \beta_i^f, \quad i = l_1+1, \dots, l.$$

The solution exists iff

$$P^f := \left\{ \frac{1 - \beta_i^f \bar{\beta}_j^f}{\alpha_i^f + \bar{\alpha}_j^f} \right\} > 0, \quad i = l_1+1, \dots, l,$$

$$j = l_1+1, \dots, l$$

and is given by

$$u_j^f(p) := \frac{(p - \alpha_j^f)u_{j+1}^f(p) + \rho_j^f(p + \bar{\alpha}_j^f)}{(p + \bar{\alpha}_j^f) + \bar{\rho}_j^f(p - \alpha_j^f)u_{j+1}^f(p)},$$

$$j = l, \dots, l_1 + 1, \quad (3)$$

$$u^f(p) := u_{l_1+1}^f(p) \quad (4)$$

where $u_{l_1+1}^f(p)$ is any SBR function and

$$\beta_{i,1}^f := \beta_i^f, \quad i = l_1+1, \dots, l,$$

$$\beta_{i,j+1}^f := \frac{(\alpha_i^f + \bar{\alpha}_j^f)(\beta_{i,j}^f - \beta_{j,j}^f)}{(\alpha_i^f - \alpha_j^f)(1 - \bar{\beta}_{j,j}^f \beta_{i,j}^f)},$$

$$l_1+1 \leq j \leq i-1 \leq l-1,$$

$$\rho_j^f := \beta_{j,j}^f.$$

$P^f > 0$ iff $|\beta_{i,j}^f| < 1$. Hence, $P^f > 0$ implies $|\rho_j^f| < 1$.

2) The Slow NP Problem: Consider two sequences of l_1 complex numbers

$$\alpha_1^s, \dots, \alpha_{l_1}^s$$

$$\beta_1^s, \dots, \beta_{l_1}^s$$

with the property that $\operatorname{Re}[\alpha_i^s] > 0$ and $|\beta_i^s| < 1$ for $i = 1, \dots, l_1$. We seek to find an SBR function $u^s(s)$ such that

$$u^s(\alpha_i^s) = \beta_i^s, \quad i = 1, \dots, l_1.$$

The solution exists iff

$$P^s := \left\{ \frac{1 - \beta_i^s \bar{\beta}_j^s}{\alpha_i^s + \bar{\alpha}_j^s} \right\} > 0, \quad i = 1, \dots, l_1, \quad j = 1, \dots, l_1$$

and is given by

$$u_j^s(s) := \frac{(s - \alpha_j^s)u_{j+1}^s(s) + \rho_j^s(s + \bar{\alpha}_j^s)}{(s + \bar{\alpha}_j^s) + \bar{\rho}_j^s(s - \alpha_j^s)u_{j+1}^s(s)}, \quad j = l_1, \dots, 1, \quad (5)$$

$$u^s(s) := u_{l_1+1}^s(s) \quad (6)$$

where $u_{l_1+1}^s(s)$ is any SBR function and

$$\beta_{i,1}^s := \beta_i^s, \quad i = 1, \dots, l_1,$$

$$\beta_{i,j+1}^s := \frac{(\alpha_i^s + \bar{\alpha}_j^s)(\beta_{i,j}^s - \beta_{j,j}^s)}{(\alpha_i^s - \alpha_j^s)(1 - \bar{\beta}_{j,j}^s \beta_{i,j}^s)}, \quad 1 \leq j \leq i-1 \leq l_1-1,$$

$$\rho_j^s := \beta_{j,j}^s.$$

$P^s > 0$ iff $|\beta_{i,j}^s| < 1$. Hence, $P^s > 0$ implies $|\rho_j^s| < 1$.

We relate the slow and fast NP problems to the full-order NP problem by setting

$$(\alpha_j^s, \beta_j^s) := (\alpha_j(0), \beta_j(0)), \quad j = 1, \dots, l_1,$$

$$(\alpha_j^f, \beta_j^f) := (\alpha_j(0), \beta_j(0)), \quad j = l_1+1, \dots, l.$$

It then follows that $\beta_{i,j}^s = \beta_{i,j}(0)$, $\rho_j^s = \rho_j(0)$ for $i, j = 1, \dots, l_1$, and $\beta_{i,j}^f = \beta_{i,j}(0)$, $\rho_j^f = \rho_j(0)$ for $i, j = l_1+1, \dots, l$.

Our objectives are as follows:

- to show that the solvability condition for the full-order problem is implied by the solvability conditions for the slow and the fast subproblems;
- to construct an SBR-TFS transfer function $u^{s,f}(s, \epsilon)$ from the solutions of the slow and fast subproblems which approximates the solution $u(s, \epsilon)$ of the full-order problem arbitrarily closely, i.e., an SBR transfer function $u^{s,f}(s, \epsilon) \in T_\epsilon$ such that

$$\|u(s, \epsilon) - u^{s,f}(s, \epsilon)\|_\infty = O(\epsilon).$$

Since the size of data are smaller for the slow and the fast problems when compared to the full-order problem, it is evident that one gains a computational advantage by solving for $u^{s,f}(s, \epsilon)$ instead of $u(s, \epsilon)$.

III. ON THE SOLVABILITY CONDITION

In this section, we will show that the NP problem is solvable for the full-order problem if it is solvable for the slow and fast subproblems.

Theorem 1: Let $P^s > 0$ and $P^f > 0$. Then there exists an $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*)$, $P(\epsilon) > 0$.

Proof Since $P_{11}(0) = P^* > 0$, for sufficiently small ϵ $P_{11}(\epsilon) > 0$. Hence, for sufficiently small ϵ , $P_{11}^{-1}(\epsilon)$ exists. Thus, using a well-known result on the similarity transformation [11] we conclude that $P(\epsilon)$ is similar to the lower block triangular matrix

$$\begin{bmatrix} P_{11}(\epsilon) & 0 \\ \star & \epsilon[P_{22}(\epsilon) - \epsilon P_{12}(\epsilon)P_{11}^{-1}(\epsilon)\bar{P}_{12}(\epsilon)] \end{bmatrix}$$

where \star stands for the block whose value is unimportant in the following development. Therefore $P(\epsilon) > 0$ iff $L(\epsilon) = P_{22}(\epsilon) - \epsilon P_{12}(\epsilon)P_{11}^{-1}(\epsilon)\bar{P}_{12}(\epsilon) > 0$. But since $L(0) = P_{22}(0) = P^* > 0$ we see that for sufficiently small ϵ $L(\epsilon) > 0$. Hence there exists an $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*)$, conditions $P^* > 0$ and $P^* > 0$ imply $P(\epsilon) > 0$. \square F. D.

IV. THE TFS SOLUTION

The main step towards the construction of $u^l(s, \epsilon)$ is to find conditions which ensure a TFS solution to the full order algorithm (1)–(2). To this end we introduce the following family of transfer functions

$$u_j^s(s) = \frac{-\alpha_j^l u_{j+1}^s(s) + \rho_j^l \alpha_j^l}{\bar{\alpha}_j^l - \bar{\rho}_j^l \alpha_j^l u_{j+1}^s(s)} \quad j = l, \dots, l_1 + 1 \quad (7)$$

$$u_j^l(s) = \frac{u_{j+1}^l(p) + \rho_j}{1 + \rho_j^s u_{j+1}^l(p)} \quad j = l_1, \dots, 1 \quad (8)$$

and prove a lemma about them.

Lemma 2 Let $u_{j+1}^s(s)$ and $u_{j+1}^l(p)$ be two SBR transfer functions. Then $u_j^s(s)$ defined by (7) and $u_j^l(p)$ defined by (8) are SBR transfer functions for $j = l, \dots, l_1 + 1$ and $j = l_1, \dots, 1$ respectively.

Proof We prove by induction that $u^l(p)$ is an SBR transfer function. The proof for $u_j(s)$ is quite similar and is omitted. We are given that $u_{l_1+1}^l(p)$ is an SBR transfer function. So suppose $u_{j+1}^l(p)$ is also an SBR transfer function and consider $u_j^l(p)$ obtained from (8). Certainly $u_j^l(p)$ is a proper rational function. It is also analytic in the closed right half plane since otherwise $1 + \rho_j u_{j+1}^l(\zeta) = 0$ for some ζ with $\text{Re}[\zeta] > 0$. But then $|u_{j+1}^l(\zeta)| = 1/|\rho_j^s| > 1$ which leads to a contradiction in view of the maximum principle. To complete the proof we have by a simple calculation

$$1 - |u_j^l(i\omega)|^2 = \frac{(1 - |\rho_j^s|)(1 - |u_{j+1}^l(i\omega)|^2)}{|1 + \rho_j u_{j+1}^l(i\omega)|^2}$$

Since $|\rho_j| < 1$ and $|u_{j+1}^l(i\omega)| < 1$ it follows that $u_j^l(p)$ is an SBR function. \square F. D.

We now use Lemma 2 to show that the algorithm (1)–(2) will indeed produce a TFS solution provided it is initialized with a correct TFS transfer function.

Theorem 2 Let $u_{l_1+1}^l(p)$ and $u_{l_1+1}^s(s)$ be two SBR transfer functions used to initiate the algorithms (3)–(4) and (5)–(6) respectively where $u_{l_1+1}^s(s)$ is generated from (7) using an SBR transfer function $u_{l_1+1}^l(p)$ as the initial data. Let $u_{l_1+1}(s, \epsilon) \in \mathcal{I}$ have no unstable lost poles and let $u_{l_1+1}^s(s)$ and $u_{l_1+1}^l(p)$ be its slow and fast transfer functions, respectively. Then the transfer function $u(s, \epsilon)$ generated via the full order algorithm (1)–(2) is an SBR-TFS transfer function.

Proof By Lemma 1 $u_{l_1+1}(s, \epsilon)$ is an SBR transfer function. Therefore from the general theory $u(s, \epsilon)$ produced by algorithm (1)–(2) is also an SBR transfer function. It remains to show that $u(s, \epsilon)$ is a TFS transfer function. For that, we consider two cases.

Case I $l_1 + 1 \leq j \leq l$. We prove by induction. By assumption $u_{l_1+1}(s, \epsilon) \in \mathcal{I}$. So suppose $u_{j+1}(s, \epsilon) \in \mathcal{I}$ and invoke [9 Theorem 2.2] to write

$$u_{j+1}(s, \epsilon) = \frac{n_{j+1}(s, \epsilon) + s^{k_{j+1}} n_{2,j+1}(s, \epsilon)}{d_{j+1}(s, \epsilon) + s^{k_{j+1}} d_{2,j+1}(s, \epsilon)} \quad (9)$$

where

- 1) $n_{j+1}(s, \epsilon)$, $n_{2,j+1}(s, \epsilon)$, $d_{j+1}(s, \epsilon)$ and $d_{2,j+1}(s, \epsilon)$ are polynomials with coefficients analytic at $\epsilon = 0$.
- 2) $\deg[d_{j+1}(s, \epsilon)] = \deg[d_{j+1}(s, 0)] - l$.
- 3) $\deg[d_{2,j+1}(p, \epsilon)] = \deg[d_{2,j+1}(p, 0)]$.
- 4) $\deg[n_{j+1}(s, \epsilon)] \leq k_{j+1}$.
- 5) $\deg[n_{2,j+1}(p, \epsilon)] \leq \deg[d_{2,j+1}(p, \epsilon)]$.
- 6) The constant terms of $n_{j+1}(p, \epsilon)$ and $d_{2,j+1}(p, \epsilon)$ are both zero.

Now the first and the last conditions allow us to expand the polynomials in (9) as

$$n_{j+1}(s, \epsilon) = \sum_{k=0}^{k_{j+1}} \Lambda_{j+1}^k(\epsilon) s^k$$

$$n_{2,j+1}(s, \epsilon) = \sum_{k=1}^{k_{j+1}} \Lambda_{j+1}^k(\epsilon) s^k$$

$$d_{j+1}(s, \epsilon) = \sum_{k=0}^{k_{j+1}} D_{j+1}^k(\epsilon) s^k$$

$$d_{2,j+1}(s, \epsilon) = \sum_{k=1}^{k_{j+1}} D_{j+1}^k(\epsilon) s^k$$

So using these expansions we substitute (9) in (1) and express $u_j(s, \epsilon)$ as

$$u_j(s, \epsilon) = \frac{n_{j+1}(s, \epsilon) + s^{k_{j+1}} n_{2,j+1}(s, \epsilon)}{d_{j+1}(s, \epsilon) + s^{k_{j+1}} d_{2,j+1}(s, \epsilon)} \quad (10)$$

where $k_j = k_{j+1}$ and

$$\begin{aligned} n_{j+1}(s, \epsilon) &= (\epsilon s - \alpha_j(\epsilon)) n_{j+1}(s, \epsilon) \\ &\quad + \rho_j(\epsilon)(\epsilon s + \bar{\alpha}_j(\epsilon)) d_{j+1}(s, \epsilon) \\ &\quad - \epsilon s (\Lambda_{j+1}^{k_{j+1}}(\epsilon) + \rho_j(\epsilon) D_{j+1}^{k_{j+1}}(\epsilon)) s^{k_{j+1}-1} \end{aligned} \quad (11)$$

$$\begin{aligned} n_{j+1}(\epsilon s, \epsilon) &= (\epsilon s - \alpha_j(\epsilon)) n_{j+1}(\epsilon s, \epsilon) + \rho_j(\epsilon)(\epsilon s + \bar{\alpha}_j(\epsilon)) \\ &\quad d_{j+1}(\epsilon s, \epsilon) + \epsilon s (\Lambda_{j+1}^{k_{j+1}}(\epsilon) + \rho_j(\epsilon) D_{j+1}^{k_{j+1}}(\epsilon)) \end{aligned} \quad (12)$$

$$\begin{aligned} d_{j+1}(s, \epsilon) &= (\epsilon s + \bar{\alpha}_j(\epsilon)) d_{j+1}(s, \epsilon) + \bar{\rho}_j(\epsilon)(\epsilon s - \alpha_j(\epsilon)) \\ &\quad n_{j+1}(s, \epsilon) - \epsilon s (D_{j+1}^{k_{j+1}}(\epsilon) + \bar{\rho}_j(\epsilon) \Lambda_{j+1}^{k_{j+1}}(\epsilon)) s^{k_{j+1}-1} \end{aligned} \quad (13)$$

$$\begin{aligned} d_{2,j+1}(\epsilon s, \epsilon) &= (\epsilon s + \bar{\alpha}_j(\epsilon)) d_{2,j+1}(\epsilon s, \epsilon) + \bar{\rho}_j(\epsilon)(\epsilon s - \alpha_j(\epsilon)) \\ &\quad n_{2,j+1}(\epsilon s, \epsilon) + \epsilon s (D_{j+1}^{k_{j+1}}(\epsilon) + \bar{\rho}_j(\epsilon) \Lambda_{j+1}^{k_{j+1}}(\epsilon)) \end{aligned} \quad (14)$$

We claim that

- 1) $n_{j+1}(s, \epsilon)$, $n_{2,j+1}(s, \epsilon)$, $d_{j+1}(s, \epsilon)$ and $d_{2,j+1}(s, \epsilon)$ are polynomials with coefficients analytic at $\epsilon = 0$. This follows since all polynomials in (9) have coefficients analytic at $\epsilon = 0$.
- 2) $\deg[d_{j+1}(s, \epsilon)] = \deg[d_{j+1}(s, 0)] = k_j$. To prove this claim we note that if

$$\bar{\alpha}_j^l D_{j+1}^{k_{j+1}}(0) - \bar{\rho}_j^l \alpha_j^l \Lambda_{j+1}^{k_{j+1}}(0) \neq 0 \quad (15)$$

then degree of $d_{j+1}(s, \epsilon)$ in (13) does not drop to a lower value when ϵ is set to zero. That is, for sufficiently small ϵ , the claim is true. On the other hand, setting $\epsilon = 0$ in (1) and comparing the result with (7) we see by induction that $u_j(s, 0) = u_j(s)$

for $j = l, \dots, l_1 + 1$. Thus, upon computing $u_{j+1}^*(s)$ from (9) and substituting the result in (7), we deduce that

$$u_j^*(\infty) = \frac{-\alpha_j^l N_{1,j+1}^{k'+1}(0) + \rho_j^l \bar{\alpha}_j^l D_{1,j+1}^{k'+1}(0)}{\bar{\alpha}_j^l D_{1,j+1}^{k'+1}(0) - \bar{\rho}_j^l \alpha_j^l N_{1,j+1}^{k'+1}(0)}.$$

But then the properness of $u_j^*(s)$ implies that $u_j^*(\infty) \neq \infty$. Hence, (15) holds.

- 3) $\deg[d_{2,j}(p, \epsilon)] = \deg[d_{2,j}(p, 0)]$. The method of proof for this part is similar to the previous part. That is, we see from (14) that if

$$D_{2,j+1}^{k'+1}(0) + \bar{\rho}_j^l N_{2,j+1}^{k'+1}(0) \neq 0 \quad (16)$$

then for sufficiently small ϵ , the claim is true. On the other hand, computing $u_j(p/\epsilon, \epsilon)|_{\epsilon=0}$ from (1) and comparing it with $u_j^l(p)$ in (3), it follows by induction that the two transfer functions are equal for each $j = l, \dots, l_1 + 1$. Thus, upon computing $u_{j+1}^l(p)$ from (9) and substituting the result in (3), we deduce that

$$u_j^l(\infty) = \frac{N_{2,j+1}^{k'+1}(0) + \rho_j^l D_{2,j+1}^{k'+1}(0)}{D_{2,j+1}^{k'+1}(0) + \bar{\rho}_j^l N_{2,j+1}^{k'+1}(0)}.$$

But the properness of $u_j^l(p)$ implies that $u_j^l(\infty) \neq 0$. Hence, (16) holds.

- 4) $\deg[u_{1,j}(s, \epsilon)] \leq k_j$. This follows easily by examining (11).
 5) $\deg[u_{2,j}(p, \epsilon)] \leq \deg[d_{2,j}(p, \epsilon)]$. This follows easily by examining (12).
 6) The constant terms of $u_{2,j}(p, \epsilon)$ and $d_{2,j}(p, \epsilon)$ are both zero. This follows by examining (12) and (14), noting that the constant terms of the polynomials $u_{2,j+1}(p, \epsilon)$ and $d_{2,j+1}(p, \epsilon)$ are zeros.

Therefore, by [9, Theorem 2.2], $u_j(s, \epsilon) \in T$, for $j = l_1 + 1, \dots, l$.

Case 2— $1 \leq j \leq l_1$. We prove by induction. By the previous step, we know that $u_{l_1+1}(s, \epsilon) \in T$. So, let us suppose $u_{j+1}(s, \epsilon) \in T$, and invoke [9, Theorem 2.2] to express $u_{j+1}(s, \epsilon)$ as in (9) with the six properties as was stated in Case 1. Next, we substitute (9) in (2) and express $u_j(s, \epsilon)$ as in (10) where $k_j = k_{j+1} + 1$, and

$$\begin{aligned} u_{1,j}(s, \epsilon) &= (s - \alpha_j(\epsilon))n_{1,j+1}(s, \epsilon) \\ &\quad + \rho_j(\epsilon)(s + \bar{\alpha}_j(\epsilon))d_{1,j+1}(s, \epsilon) \\ &\quad + \epsilon(-\alpha_j(\epsilon)N_{2,j+1}^1(\epsilon) + \rho_j(\epsilon)\bar{\alpha}_j(\epsilon)D_{2,j+1}^1(\epsilon)) \\ &\quad \cdot s^{k_{j+1}+1}. \end{aligned} \quad (17)$$

$$\begin{aligned} u_{2,j}(\epsilon s, \epsilon) &= u_{2,j+1}(\epsilon s, \epsilon) + \rho_j(\epsilon)d_{2,j+1}(\epsilon s, \epsilon) \\ &\quad - \alpha_j(\epsilon) \left(\frac{u_{2,j+1}(\epsilon s, \epsilon)}{s} - \epsilon N_{2,j+1}^1(\epsilon) \right) \\ &\quad + \rho_j(\epsilon)\bar{\alpha}_j(\epsilon) \left(\frac{d_{2,j+1}(\epsilon s, \epsilon)}{s} - \epsilon D_{2,j+1}^1(\epsilon) \right). \end{aligned} \quad (18)$$

$$\begin{aligned} d_{1,j}(s, \epsilon) &= (s + \bar{\alpha}_j(\epsilon))d_{1,j+1}(s, \epsilon) \\ &\quad + \bar{\rho}_j(\epsilon)(s - \alpha_j(\epsilon))n_{1,j+1}(s, \epsilon) \\ &\quad + \epsilon(\bar{\alpha}_j(\epsilon)D_{2,j+1}^1(\epsilon) - \bar{\rho}_j(\epsilon)\alpha_j(\epsilon)N_{2,j+1}^1(\epsilon)) \\ &\quad \cdot s^{k_{j+1}+1}. \end{aligned} \quad (19)$$

$$\begin{aligned} d_{2,j}(\epsilon s, \epsilon) &= d_{2,j+1}(\epsilon s, \epsilon) + \bar{\rho}_j(\epsilon)u_{2,j+1}(\epsilon s, \epsilon) \\ &\quad + \bar{\alpha}_j(\epsilon) \left(\frac{d_{2,j+1}(\epsilon s, \epsilon)}{s} - \epsilon D_{2,j+1}^1(\epsilon) \right) \\ &\quad - \bar{\rho}_j(\epsilon)\alpha_j(\epsilon) \left(\frac{u_{2,j+1}(\epsilon s, \epsilon)}{s} - \epsilon N_{2,j+1}^1(\epsilon) \right). \end{aligned} \quad (20)$$

We claim the following:

- 1) $u_{1,j}(s, \epsilon)$, $u_{2,j}(\epsilon s, \epsilon)$, $d_{1,j}(s, \epsilon)$, and $d_{2,j}(\epsilon s, \epsilon)$ are polynomials with coefficients analytic at $\epsilon = 0$. This follows since all polynomials in (9) have coefficients analytic at $\epsilon = 0$.
- 2) $\deg[d_{1,j}(s, \epsilon)] = \deg[d_{1,j}(s, 0)] = k_j$. This is proved by noting from (19) that if

$$D_{1,j+1}^{k'+1}(0) + \bar{\rho}_j^l N_{1,j+1}^{k'+1}(0) \neq 0 \quad (21)$$

then for sufficiently small ϵ , the claim is true. On the other hand, setting $\epsilon = 0$ in (2) and comparing the result with (5), we see by induction that $u_j(s, 0) = u_j^*(s)$ for $j = l_1, \dots, 1$. Thus, upon computing $u_{j+1}^*(s)$ from (9) and substituting the result in (5), we deduce that

$$u_j^*(\infty) = \frac{N_{1,j+1}^{k'+1}(0) + \rho_j^l D_{1,j+1}^{k'+1}(0)}{D_{1,j+1}^{k'+1}(0) + \bar{\rho}_j^l N_{1,j+1}^{k'+1}(0)}.$$

But then the properness of $u_j^*(s)$ implies that $u_j^*(\infty) \neq \infty$. Hence, (21) holds.

- 3) $\deg[d_{2,j}(p, \epsilon)] = \deg[d_{2,j}(p, 0)]$. To prove this, note from (20) that if

$$D_{2,j+1}^{k'+1}(0) + \bar{\rho}_j^l N_{2,j+1}^{k'+1}(0) \neq 0 \quad (22)$$

then for sufficiently small ϵ , the claim is true. On the other hand, computing $u_j(p/\epsilon, \epsilon)|_{\epsilon=0}$ from (2) and comparing it with $u_j^l(p)$ in (3), we see by induction that the two transfer functions are equal for each $j = l_1, \dots, 1$. Thus, upon computing $u_{j+1}^l(p)$ from (9) and substituting the result in (8), we deduce that

$$u_j^l(\infty) = \frac{N_{2,j+1}^{k'+1}(0) + \rho_j^l D_{2,j+1}^{k'+1}(0)}{D_{2,j+1}^{k'+1}(0) + \bar{\rho}_j^l N_{2,j+1}^{k'+1}(0)}.$$

But then properness of $u_j^l(s)$ implies that $u_j^l(\infty) \neq \infty$. Hence, (22) holds.

- 4) $\deg[u_{1,j}(s, \epsilon)] \leq k_j$. This follows easily by examining (11).
- 5) $\deg[u_{2,j}(p, \epsilon)] \leq \deg[d_{2,j}(p, \epsilon)]$. This follows easily by examining (12).
- 6) The constant terms of $u_{2,j}(p, \epsilon)$ and $d_{2,j}(p, \epsilon)$ are both zero. This follows by examining (12) and (14) and noting that the constant polynomials $u_{2,j+1}(p, \epsilon)$ and $d_{2,j+1}(p, \epsilon)$ are zero.

Therefore, by [9, Theorem 2.2], $u_j(s, \epsilon) \in T$, for $j = 1, \dots, l_1$. Combining the two cases, we conclude that $u_j(s, \epsilon) \in T$, for all j . This completes the proof.

Q.E.D.

We now present an algorithm for the construction of the TFS-SBR transfer function $u^{s^k}(s, \epsilon)$.

Algorithm.

- 1) Check to see if $P^l > 0$ and $P^h > 0$. If these conditions are satisfied, proceed to the next step. Otherwise, terminate the algorithm.
- 2) Choose an SBR transfer function $u_{l_1+1}^l(p)$ and compute $u_{l_1+1}^F(p)$ from (3).
- 3) Choose an SBR transfer function $u_{l_1+1}^*(s)$ and compute $u_{l_1+1}^*(s)$ from (7).
- 4) Use the SBR transfer function $u_{l_1+1}^F(p)$ obtained in Step 2 to compute $u_1^F(p)$ from (8).
- 5) Use the SBR transfer function $u_{l_1+1}^*(s)$ obtained in Step 3 to compute $u_1^*(s)$ from (5).
- 6) Set $u^l(p) := u_1^l(p)$ and $u^h(s) := u_1^*(s)$.
- 7) Set $u^{s^k}(s, \epsilon) := u^h(s) + u^l(\epsilon s) - u^h(\infty)$.

We have the following theorem.

Theorem 3 Let $u(s, \epsilon)$ be the solution to the full-order NP problem (1)–(2) with the initial data given by Theorem 2 and let $u^{s,f}(s, \epsilon)$ be the transfer function constructed in the aforementioned algorithm. Then for sufficiently small ϵ

- 1) $u^{s,f}(s, \epsilon)$ is a TFS–SBR transfer function, and
- 2) $\|u(s, \epsilon) - u^{s,f}(s, \epsilon)\|_{\infty} = O(\epsilon)$

Proof By Theorem 2, $u(s, \epsilon)$ is a TFS–SBR transfer function. Also $u^{s,f}(s, \epsilon) \in T$ because it has the same slow and fast transfer functions as $u(s, \epsilon)$. Since both of these slow and fast transfer functions are SBR functions by Lemma 1, $u^{s,f}(s, \epsilon)$ is also an SBR transfer function. This proves part 1 of the theorem. Part 2 follows from [9, Theorem 4.1].

QED

V CONCLUSION

In this note we have considered a version of the NP interpolation problem. We have obtained conditions which guarantee a TFS solution. We have formulated two smaller NP interpolation problems and have shown that they can be solved in parallel. We have reduced the solvability condition in terms of the solvability conditions for these two smaller problems. The immediate gain here is the reduction in the verification process and hence the computer time. We have used the solutions obtained from these two smaller problems and have constructed a solution which can become arbitrarily close in the H_{∞} norm to the solution of the original NP problem. The results of this note should prove useful both from the computational standpoint as well as in the areas such as robust stabilization problem for TFS systems.

REFERENCES

- [1] P. Delsarte, Y. Genin, and Y. Kump, "On the role of the Nevanlinna–Pick problem in circuit and system theory," *Int. J. Circuit Theory Appl.*, vol. 9, pp. 177–187, 1981.
- [2] B. A. Francis and G. Zames, "On the optimal sensitivity theory for SISO feedback systems," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 9–16, 1984.
- [3] B. C. Chang and J. J. B. Pearson, "Optimal disturbance reduction in linear multivariable systems," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 880–887, 1984.
- [4] H. Kimura, "Robust stabilizability for a class of transfer functions," *IEEE Trans. Automat. Contr.*, vol. AC-29, no. 9, pp. 788–793, 1984.
- [5] P. P. Khargonekar and A. Tannenbaum, "Non-Euclidean metrics and the robust stabilization of systems with parameter uncertainty," *IEEE Trans. Automat. Contr.*, vol. AC-30, no. 10, pp. 1005–1013, 1985.
- [6] N. I. Akhiezer, *The Classical Moment Problem*. London, U.K.: Oliver and Boyd, 1965.
- [7] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, 4th ed. AMS Colloquium Publications, 1965.
- [8] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum, *Feedback Control Theory*. New York: Macmillan, 1992.
- [9] D. W. Luse and H. K. Khalil, "Frequency domain results for systems with slow and fast dynamics," *IEEE Trans. Automat. Contr.*, vol. AC-30, no. 12, pp. 1171–1179, 1985.
- [10] D. W. Luse and J. A. Ball, "Frequency scale decomposition of H_{∞} disk problems," *SIAM J. Contr. Optim.*, vol. 27, pp. 814–835, 1989.
- [11] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 2nd ed. Baltimore, MD: Johns Hopkins Univ., 1990.

Multiscale Smoothing Error Model

Mark R. Luetgten and Alan S. Willsky

Abstract—A class of multiscale stochastic models used in recursive dynamics on trees has recently been introduced. These models are interesting because they can be used to represent a wide class of physical phenomena and because they lead to efficient algorithms for estimation and likelihood calculation. In this paper, we provide a complete statistical characterization of the error associated with smoothed estimates of the multiscale stochastic processes described by these models. In particular, we show that the smoothing error is itself a multiscale stochastic process with parameters that can be explicitly calculated.

I INTRODUCTION

A class of multiscale models describing stochastic processes indexed by the nodes of a tree has recently been introduced in [1]–[2]. These models can be used to capture a surprisingly rich class of physical phenomena. For instance, experimental results in [2] illustrate that they can be used to model the statistical self-similarity exhibited by stochastic processes with generalized power spectra of the form $1/f^{\alpha}$ and in [3] we describe how they can be used to represent any 1-D Markov process or 2-D Markov random field. Moreover, this class of models leads to efficient algorithms for estimation and likelihood calculation and as a result provides a useful framework for a variety of signal and image processing problems [1]–[2], [4]–[6].

Knowledge of the error statistics of smoothed estimates of such processes is essential for the development of a number of important new applications, including for instance so-called mapping problems [7], the multiscale counterpart to the model validation problem in [8], and certain oceanographic problems [9]. Several such applications have been developed in the context of 1-D Gauss–Markov models by exploiting relatively recent results that show that the smoothing error processes associated with Gauss–Markov models are themselves Gauss–Markov processes [7], [8], [10], [11].¹ In this paper, we derive a dynamic model for the smoothing error process associated with multiscale stochastic models. In particular, we show that the smoothing error is itself a multiscale stochastic process with parameters that can be explicitly computed. These results generalize previous results for Gauss–Markov processes, since these processes correspond to a degenerate form of the multiscale models, and provide the necessary framework for applications such as those mentioned above.

This paper is organized as follows. In Section II we briefly review the class of multiscale stochastic models of interest here and the scale recursive estimation algorithm associated with them. In Section III we derive a multiscale model for the smoothing error process.

Manuscript received October 19, 1993; revised March 28, 1994. This work was supported by the Air Force Office of Scientific Research under Grant AFOSR-2-J-0002, by the Army Research Office under Grant DAAL03-92-G-0115, by the Office of Naval Research under Grant N00014-91-J-1004, and by the Advanced Research Projects Agency through Air Force Grant F49620-93-1-0604.

M. R. Luetgten is with Alphatech Inc., Burlington, MA 01803, USA. A. S. Willsky is with MIT Laboratory for Information and Decision Systems, Cambridge, MA 02139, USA.
IEEE Log Number 9406982.

¹More generally, Levy *et al.* [12] have recently shown that the smoothing error processes associated with the class of Gaussian reciprocal processes, which contains the class of Gauss–Markov processes, are themselves Gaussian reciprocal. See also [13] for similar results corresponding to 2-D Gauss–Markov random fields.

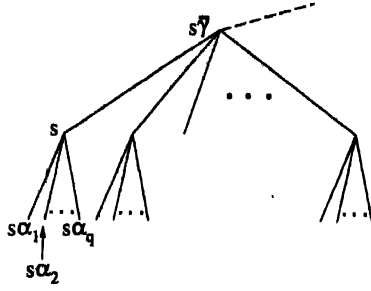


Fig. 1. Multiscale stochastic processes are indexed by a q th-order tree. The parent of a node s on the tree is denoted $s\bar{\gamma}$, and its q offspring are denoted $s\alpha_1, \dots, s\alpha_q$.

II. MULTISCALE STOCHASTIC MODELING AND OPTIMAL ESTIMATION

The models presented in this section describe multiscale Gaussian stochastic processes indexed by nodes on a tree. A q th order tree is a pyramidal structure of nodes connected such that each node of the tree has q offspring (see Fig. 1). We denote nodes on the tree with an abstract index s , and define an upward (fine-to-coarse) shift operator $\bar{\gamma}$ such that $s\bar{\gamma}$ is the parent of node s . We also define a corresponding set of downward shift operators $\alpha_1, \dots, \alpha_q$ such that $s\alpha_1, \dots, s\alpha_q$ are the offspring of node s . In addition, we denote the set of nodes on the tree as \mathcal{T} and the set of nodes that includes node s and all of its descendants as \mathcal{T}_s , i.e., $\mathcal{T}_s = \{\sigma | \sigma = s \text{ or } \sigma \text{ is a descendant of } s\}$. Also, the complement of \mathcal{T}_s is denoted \mathcal{T}_s^c . The statistical characterization of the model state $x(s) \in \mathcal{R}^n$ is then given by

$$x(s) = A(s)x(s\bar{\gamma}) + B(s)w(s) \quad (1)$$

under the assumptions that $x(0) \sim \mathcal{N}(0, P(0))$, $w(s) \sim \mathcal{N}(0, I)$, $A(s)$ and $B(s)$ are matrices of appropriate size, and $s = 0$ is the root node at the top of the tree. The driving noise $w(s) \in \mathcal{R}^m$ is white, i.e., $w(s)$ and $w(\sigma)$ are independent if $s \neq \sigma$, and independent of the initial condition $x(0)$. The class of models (1) has a statistical structure that can be exploited to develop efficient signal processing algorithms. In particular, note that any given node on the q th-order tree can be viewed as a boundary between $q + 1$ subsets of nodes (q corresponding to paths leading towards offspring and one corresponding to a path leading towards a parent). An important property of the model (1) is that, conditioned on the value of the state at any node, the values of the state corresponding to the $q + 1$ subsets of nodes are independent. This fact is the basis for the development in [1], [2] of an algorithm for computing smoothed estimates of $x(s)$ based on noisy measurements $y(s) \in \mathcal{R}^p$ of the form

$$y(s) = C(s)x(s) + v(s) \quad (2)$$

where $v(s) \sim \mathcal{N}(0, R(s))$, and is independent of both $w(s)$ and $x(0)$. The algorithm for computing the smoothed estimates of $x(s)$ is a generalization to q th-order trees of the well-known Rauch–Tung–Striebel algorithm for smoothing 1-D Gauss–Markov processes. We briefly review the multiscale smoothing algorithm next, and then derive a model for the error associated with the smoothed estimates.

We denote the set of states defined at nodes in \mathcal{T}_s as X_s , i.e., $X_s = \{x(\sigma) | \sigma \in \mathcal{T}_s\}$, and similarly $Y_s = \{y(\sigma) | \sigma \in \mathcal{T}_s\}$. The set of measurements in the subtree strictly below s is denoted Y_s^+ , i.e., $Y_s^+ = \{y(\sigma) | \sigma \text{ is a descendant of } s\}$. We also define $\hat{x}(s|Y)$ as the expected value of $x(s)$ given measurements in the set Y and the corresponding error covariance as $P(s|Y)$.

The upward sweep of the smoothing algorithm begins with the initialization of $\hat{x}(s|Y_s^+)$ and $P(s|Y_s^+)$ at the finest level. In particular,

for every s at this finest scale we set $\hat{x}(s|Y_s^+)$ to zero and $P(s|Y_s^+)$ to the solution at the finest level of the tree of the Lyapunov equation:

$$P(s) = A(s)P(s\bar{\gamma})A^T(s) + B(s)B^T(s) \quad (3)$$

where $P(s)$ denotes the covariance of the process $x(s)$ at node s . Suppose then that we have $\hat{x}(s|Y_s^+)$ and $P(s|Y_s^+)$ at a given node s . This estimate is *updated* to incorporate the measurement $y(s)$ according to the following:

$$\hat{x}(s|Y_s) = \hat{x}(s|Y_s^+) + K(s)[y(s) - C(s)\hat{x}(s|Y_s^+)] \quad (4)$$

$$P(s|Y_s) = [I - K(s)C(s)]P(s|Y_s^+) \quad (5)$$

where $K(s) = P(s|Y_s^+)C^T(s)[C(s)P(s|Y_s^+)C^T(s) + R(s)]^{-1}$.

Suppose next that we have the updated estimates $\hat{x}(s\alpha_i|Y_{s\alpha_i})$ at all of the immediate descendants of node s . The next step involves the use of these estimates to predict $x(s)$ at the next coarser scale, i.e., to compute $\hat{x}(s|Y_{s\alpha_i})$. Using the following *upward* model for the multiscale process [1], [2]:

$$x(s\bar{\gamma}) = F(s)x(s) + \bar{w}(s) \quad (6)$$

with the measurement equation again given by (2), and where $F(s) = P(s\bar{\gamma})A^T(s)P(s)^{-1}$ and $E[\bar{w}(s)\bar{w}^T(s)] = P(s\bar{\gamma}) - P(s\bar{\gamma})A^T(s)P(s)^{-1}A(s)P(s\bar{\gamma}) \equiv Q(s)$, we compute the fine-to-coarse *predicted* estimates:

$$\hat{x}(s|Y_{s\alpha_i}) = F(s\alpha_i)\hat{x}(s\alpha_i|Y_{s\alpha_i}) \quad (7)$$

$$P(s|Y_{s\alpha_i}) = F(s\alpha_i)P(s\alpha_i|Y_{s\alpha_i})F^T(s\alpha_i) + Q(s\alpha_i). \quad (8)$$

The estimates $\hat{x}(s|Y_{s\alpha_i})$, $i = 1, \dots, q$ are then *merged* to obtain

$$\hat{x}(s|Y_s^+) = P(s|Y_s^+) \sum_{i=1}^q P^{-1}(s|Y_{s\alpha_i}) \hat{x}(s|Y_{s\alpha_i}) \quad (9)$$

$$P(s|Y_s^+) = \left[(1-q)P(s)^{-1} + \sum_{i=1}^q P^{-1}(s|Y_{s\alpha_i}) \right]^{-1}. \quad (10)$$

We assume here that $P(s)$ and $P(s\bar{\gamma}|Y_s)$ are invertible for all s so that the upward model given by (6) and the merge operation given by (9), (10) are well-defined. As discussed at the end of the next section, this restriction can be removed.

The recursion given by the update, predict and merge equations proceeds up the tree until one obtains the smoothed estimate of the root node, $\hat{x}(0|Y_0)$. This estimate initializes a *downward sweep* in which $\hat{x}(s|Y_0)$ is computed according to

$$\hat{x}(s|Y_0) = \hat{x}(s|Y_s) + J(s)[\hat{x}(s\bar{\gamma}|Y_0) - \hat{x}(s\bar{\gamma}|Y_s)] \quad (11)$$

$$P(s|Y_0) = P(s|Y_s) + J(s)[P(s\bar{\gamma}|Y_0) - P(s\bar{\gamma}|Y_s)]J^T(s) \quad (12)$$

$$J(s) = P(s|Y_s)F^T(s)P^{-1}(s\bar{\gamma}|Y_s). \quad (13)$$

Note that (12) characterizes the smoothing error covariance at any given lattice site s , but does not provide information about the correlation structure of the error process. The goal in the next section is to provide a multiscale model for the smoothing error process, i.e., to show that the error satisfies a recursion of the form (1), and to calculate the associated model parameters. This then provides the complete statistical characterization of the smoothing error that we seek.

III. MULTISCALE SMOOTHING ERROR MODELS

Given two nodes s and $\sigma \in \mathcal{T}_s^c$ on the tree, we can represent $x(\sigma)$ in terms of $x(s\bar{\gamma})$ and an additive noise term $\varphi_{\sigma, s\bar{\gamma}}$:

$$x(\sigma) = \Phi_{\sigma, s\bar{\gamma}}x(s\bar{\gamma}) + \varphi_{\sigma, s\bar{\gamma}} \quad (14)$$

with $\varphi_{\sigma, s\bar{\gamma}}$ independent of the set of states $x(s\bar{\gamma}) \cup X_s$ and the corresponding set of measurements $y(s\bar{\gamma}) \cup Y_s$, by tracing a path

from σ to s^- and using the upward dynamics (6) and downward dynamics (1) to eliminate state variables along the way. The state transition matrix Φ_{σ, s^-} in this construction is a function of the upward and downward prediction matrices A and F along the path, whereas φ_{σ, s^-} is a linear function of the upward and downward driving noises u and π . Since φ_{σ, s^-} is independent of the set of states $i(s^-) \cup \mathcal{Y}_s$, and the measurements $y(s^-) \cup \mathcal{Y}$, we have that $i(\sigma|\mathcal{Y}_s) = \Phi_{\sigma, s^-} i(s^-|\mathcal{Y}_s)$ which using (14) implies that

$$i(\sigma|\mathcal{Y}_s) = \Phi_{\sigma, s^-} i(s^-|\mathcal{Y}) + \varphi_{\sigma, s^-} \quad (15)$$

where we have defined the error in $i(s|\mathcal{Y})$ as $i(s|\mathcal{Y}) \equiv i(s) - i(s|\mathcal{Y})$. As a result we see that $i(s|\mathcal{Y})$ has the Markov property

$$\begin{aligned} E\{i(s|\mathcal{Y})|i(\sigma|\mathcal{Y})\} &= E\{i(s|\mathcal{Y}_s)|i(s^-|\mathcal{Y})\} \quad \{\varphi_{\sigma, s^-} \mid s \in \mathcal{I}_s\} \\ &= E\{i(s|\mathcal{Y})|i(s^-|\mathcal{Y})\} + E\{i(s|\mathcal{Y})|\{\varphi_{\sigma, s^-} \mid s \in \mathcal{I}\}\} \\ &= E\{i(s|\mathcal{Y})|i(s^-|\mathcal{Y})\} \end{aligned} \quad (16)$$

The first equality in (16) follows from (15), the second from the orthogonality of φ_{σ, s^-} to $i(s^-)$ and \mathcal{Y} , and the last from the orthogonality of φ_{σ, s^-} to $i(s)$ and \mathcal{Y} . Now using the upward dynamics (6), the upward sweep prediction equation (7), and standard linear least squares formulas we can write

$$i(s|\mathcal{Y}) = J(s)i(s^-|\mathcal{Y}) + u(s) \quad (17)$$

where $J(s)$ is given by (13) and where from (16) $u(s)$ is independent of $\{i(\sigma|\mathcal{Y})|\sigma \in \mathcal{I}\}$ and has covariance

$$P(s|\mathcal{Y}) = P(s|\mathcal{Y})F^T(s)P^{-1}(s^-|\mathcal{Y})F(s)P(s|\mathcal{Y}) \quad (18)$$

Next note that the independence of $u(s)$ and $\{i(\sigma|\mathcal{Y})|\sigma \in \mathcal{I}\}$ implies that $u(s)$ is also independent of the residual information about $i(s)$ that is contained in the set of all available measurements \mathcal{Y}_s , but not contained in \mathcal{Y} . In particular, at each node in \mathcal{T} , a residual component $r(\sigma)$ that is orthogonal to the measurements in the set \mathcal{Y} can be defined as

$$\begin{aligned} r(\sigma) &= y(\sigma) - E\{y(\sigma)|\mathcal{Y}\} \\ &= C(\sigma)i(\sigma|\mathcal{Y}) + v(\sigma) \end{aligned} \quad (19)$$

Denoting $r = \{r(\sigma)|\sigma \in \mathcal{I}\}$ it is clear that $\text{span } \mathcal{Y} = \text{span } \{r\}$ that $r \perp \mathcal{Y}$ and that $r \perp u(s)$. Taking the expected value of both sides of (17) conditioned on r we obtain

$$E\{i(s|\mathcal{Y})|r\} = J(s)E\{i(s^-|\mathcal{Y})|r\} \quad (20)$$

Finally noting that

$$i(s|\mathcal{Y}_s) = i(s|\mathcal{Y}) + E\{i(s|\mathcal{Y})|r\} \quad (21)$$

and then subtracting (20) from (17) results in

$$i(s|\mathcal{Y}_0) = J(s)i(s^-|\mathcal{Y}_0) + u(s) \quad (22)$$

which is a multiscale model for the smoothing error of precisely the same form as (1).

This model is, of course, consistent with the error covariance computation in (12). In particular, using the Lyapunov equation for (22) we obtain

$$\begin{aligned} P(s|\mathcal{Y}_0) &= J(s)P(s^-|\mathcal{Y}_0)J^T(s) + P(s|\mathcal{Y}) \\ &= P(s|\mathcal{Y}_s)F^T(s)P^{-1}(s^-|\mathcal{Y}_s)F(s)P(s|\mathcal{Y}_s) \\ &= P(s|\mathcal{Y}_s) + J(s)[P(s^-|\mathcal{Y}_0) - P(s^-|\mathcal{Y}_s)]J^T(s) \end{aligned} \quad (23)$$

In addition, on first-order trees, the model (1) reduces to a standard Gauss-Markov model and hence (22) generalizes to q th order trees the corresponding 1-D time series result. The derivation here is

related to but is in fact substantially simpler than the based on backwards prediction error models in [8].

Finally we note that it is possible to derive a multiscale error model without assuming invertibility of $I(s^-)$. We refer the reader to Appendix D of [14] for a detailed derivation of a likelihood calculation algorithm for (1)-(2) that allows for rank deficient $P(s)$ and $P(s^-|\mathcal{Y})$. A slight variation of the technique in that derivation can be used to show that a multiscale error model allowing for rank deficient $P(s)$ and $I(s^-|\mathcal{Y})$ can be written precisely as in (22) but with the gain $J(s)$ given by

$$J(s) = P(s|\mathcal{Y}_s)F^T(s)P^\dagger(s^-|\mathcal{Y}) \quad (24)$$

$$J(s) = P(s^-|\mathcal{Y})A^T(s)P(s)^{-1} \quad (25)$$

and the covariance of $u(s)$ given by

$$P(s|\mathcal{Y}) = P(s|\mathcal{Y})F^T(s)P^\dagger(s^-|\mathcal{Y}_s)F(s)P(s|\mathcal{Y}) \quad (26)$$

where the superscript \dagger denotes the Moore-Penrose pseudo-inverse [15].

REFERENCES

- [1] K. Chou, A. Willsky, A. Benveniste, and M. Basseville, "Recursive and iterative estimation algorithms for multiscale stochastic processes," in *Proc. 28th IEEE Conf. Decision and Control*, Tampa, FL, Dec. 1989, pp. 1884-1889.
- [2] K. Chou, A. Willsky, and A. Benveniste, "Multiscale recursive estimation, data fusion and regularization," Massachusetts Inst. Technol. Lab. Inform. Decision Syst. Tech. Rep. LIDS P 2085, 1991; see also *IEEE Trans. Automat. Contr.*, Apr. 1994.
- [3] M. Luetgen, W. Karl, A. Willsky, and R. Tenney, "Multiscale representations of Markov random fields," *IEEE Trans. Signal Processing*, vol. 41, pp. 3377-3396, 1993.
- [4] M. Luetgen, W. Karl, and A. Willsky, "Efficient multiscale regularization with application to the computation of optical flow," *IEEE Trans. Image Processing*, vol. 3, pp. 41-64, 1994.
- [5] M. Luetgen and A. Willsky, "Likelihood calculation for a class of multiscale stochastic models with application to texture discrimination," Lab. Inform. Decision Syst., Massachusetts Inst. Technol. Tech. Rep. LIDS P 2186, 1993; to appear in *IEEE Trans. Image Processing*.
- [6] —, "Likelihood calculation for a class of multiscale stochastic models," in *Proc. 32nd IEEE Conf. Decision and Control*, San Antonio, TX, Dec. 1993.
- [7] M. Bello, A. Willsky, B. Levy, and D. Castanon, "Smoothing error dynamics and their use in the solution of smoothing and mapping problems," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 483-495, 1986.
- [8] M. Bello, A. Willsky, and B. Levy, "Construction and applications of discrete time smoothing error models," *Int. J. Contr.*, vol. 50, pp. 203-223, 1989.
- [9] P. Fieguth, W. Karl, and A. Willsky, "Multiresolution statistical estimation of North Pacific ocean height from Topex/Poseidon satellite altimetry," in *Proc. 1994 SPIE Conf. Neural and Stochastic Methods in Image and Signal Processing*, San Diego, CA, July 1994.
- [10] F. Badawi, A. Lindquist, and M. Pavon, "A stochastic realization approach to the smoothing problem," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 878-887, 1979.
- [11] R. Ackner and T. Kailath, "Discrete time complementary models and smoothing," *Int. J. Contr.*, vol. 49, pp. 1665-1682, 1989.
- [12] B. Levy, R. Frezza, and A. Krener, "Modeling and estimation of discrete time Gaussian reciprocal processes," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 1013-1023, 1990.
- [13] B. Levy, "Non causal estimation for Markov random fields," in *Proc. Int. Symp. MTNS 89 Vol. 1: Realization and Modeling in System Theory*, Basel, Switzerland, Birkhauser Verlag, 1990.
- [14] M. Luetgen, "Image processing with multiscale stochastic models," Ph.D. dissertation, Dep. Elec. Eng. Comput. Sci., Massachusetts Inst. Technol., May 1993.
- [15] S. Campbell and C. Meyer, *Generalized Inverses of Linear Transformations*, London, England, Pitman, 1979.

A Generalized Popov-Belevitch-Hautus Test of Observability

Bijoy K. Ghosh and Joachim Rosenthal

Abstract—In this paper, an earlier result on the problem of observability of a linear dynamical system due to Popov-Belevitch-Hautus has been generalized and applied to the problem of observing the initial condition of a linear dynamical system described on the space of d dimensional affine planes in \mathbb{R}^n .

1. INTRODUCTION AND MOTIVATION

In this paper we generalize the well known Popov-Belevitch-Hautus test (see [3]) on the observability of a linear dynamical system. Let \mathbf{K} denote either the field of real ($\mathbf{K} = \mathbb{R}$) or the field of complex ($\mathbf{K} = \mathbb{C}$) numbers. Let A be an $n \times n$ matrix and let C be a $p \times n$ matrix defined over \mathbf{K} . Consider the linear time invariant system

$$\dot{x} = Ax, \quad y = Cx, \quad x \in \mathbf{K}^n, y \in \mathbf{K}^p. \quad (1.1)$$

The well-known Hautus test [3] gives a necessary and sufficient condition, when the state vector $x(t)$ can be observed from the output measurement $y(t)$. To be precise one has the following.

Theorem 1 (Hautus [3]): System (1.1) is observable over either \mathbb{R} or \mathbb{C} if, and only if

$$\text{rank} \begin{bmatrix} A - \lambda I \\ C \end{bmatrix} = n, \quad \text{for all } \lambda \in \mathbb{C}. \quad (1.2)$$

It may be remarked that the rank can only be less than n if λ is an eigenvalue of the matrix A .

In this paper we consider dynamical systems for which the output vector is not observed exactly but can be ascertained with an ambiguity restricted to a d -dimensional affine subspace. The problem that we propose to consider is to compute if possible the initial condition and hence the states of the dynamical system up to a d -dimensional affine subspace. Thus for the dynamical system (1.1), if we assume that the output vector $y(t)$ is observed up to a d -dimensional plane given by an equation of the form

$$\theta(t)y(t) = \eta(t) \quad (1.3)$$

where $\theta(t)$ is a $(p-d) \times n$ matrix function of time having full rank for all but countably many instances of time and $\eta(t)$ is a vector function of time. The problem is to derive conditions on A and C under which $x(0)$ can be observed up to a d -dimensional plane.

The above class of problem occurs in machine vision as has already been introduced in [6], [1]. Specifically if we consider a plane in \mathbb{R}^3 with coordinates (X, Y, Z) given by

$$sZ = pX + qY + r. \quad (1.4)$$

Let us assume that the points on the plane (1.4) satisfy a dynamical system

$$\dot{\chi} = A\chi + b \quad (1.5)$$

where A is an arbitrary 3×3 matrix and b is a 3×1 vector given by

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad (1.6)$$

$$b = (b_1 \ b_2 \ b_3)^T, \quad (1.7)$$

respectively, and where χ is given by

$$\chi = (X \ Y \ Z)^T.$$

One can compute a dynamical system for the shape parameters p, q, r , and s described as follows:

$$\frac{d}{dt} \begin{pmatrix} p \\ q \\ -s \\ r \end{pmatrix} = \begin{pmatrix} -A^T & 0 \\ -b^T & 0 \end{pmatrix} \begin{pmatrix} p \\ q \\ -s \\ r \end{pmatrix}. \quad (1.8)$$

Typically a point on the plane (1.4) is observed with the aid of a CCD camera that projects (X, Y, Z) perspectively onto an image plane. Let (η_1, η_2) be the coordinates of the image plane and assume that the perspective projection is defined as

$$\eta_1 = \frac{fX}{Z+f}, \quad \eta_2 = \frac{fY}{Z+f} \quad (1.9)$$

where f is the focal length of the camera. One can easily compute a differential equation that is satisfied by the coordinates (η_1, η_2) and is given by

$$\begin{aligned} \dot{\eta}_1 &= \eta_1 + \eta_1 \eta_1 + \eta_1 \eta_2 + \frac{1}{f}(\eta_1^2 + \eta_2^2) \\ \dot{\eta}_2 &= \eta_2 + \eta_2 \eta_2 + \eta_2 \eta_1 + \frac{1}{f}(\eta_1^2 + \eta_2^2) \end{aligned} \quad (1.10)$$

where

$$\begin{pmatrix} \dot{\eta}_1 \\ \dot{\eta}_2 \\ \dot{\eta}_3 \\ \dot{\eta}_4 \\ \dot{\eta}_5 \\ \dot{\eta}_6 \\ \dot{\eta}_7 \\ \dot{\eta}_8 \\ \dot{\eta}_9 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -fb_1 & fa_{11} \\ 0 & 0 & -fb_2 & fa_{21} \\ -b'_1 & 0 & b_1 - fa_{11} & a_{11} - a_{33} \\ 0 & -b'_1 & -fa_{12} & a_{12} \\ -b'_2 & 0 & -fa_{21} & a_{21} \\ 0 & -b'_2 & b_2 - fa_{22} & a_{22} - a_{33} \\ -b'_3 & 0 & -fa_{31} & a_{31} \\ 0 & -b'_3 & -fa_{32} & a_{32} \\ 0 & 0 & -f & 1 \end{pmatrix} \begin{pmatrix} p \\ q \\ -s \\ r \end{pmatrix} \quad (1.11)$$

and where

$$b' = (b_1 - a_{13}f, b_2 - a_{23}f, b_3 - a_{33}f) \triangleq (b'_1, b'_2, b'_3). \quad (1.12)$$

The equation (1.10) is known as the "optical flow" and in the literature various algorithms exist as to how one can estimate $(\dot{\eta}_1, \dot{\eta}_2)$ for a given pair (η_1, η_2) (see Horn [4]).

For our purposes we would like to view (1.8), (1.11) as a linear system for which the output vector y is not observed but instead one observes the vector $(\dot{\eta}_1, \dot{\eta}_2, \eta_1, \eta_2)$ at various points on the image plane. Note that for almost every point on the image plane, (1.10) describes a homogeneous seven-dimensional plane in \mathbb{R}^9 . Thus if one observes $(\dot{\eta}_1, \dot{\eta}_2, \eta_1, \eta_2)$ for 3 points on the image plane, the output vector in (1.11) is observed up to a homogeneous 3-plane.

Manuscript received October 25, 1993; revised March 8, 1994 and May 14, 1994. B. K. Ghosh was supported in part by the Department of Energy under Grant DE-FG-02-90ER14140. J. Rosenthal was supported in part by the National Science Foundation under Grant DMS-9201263.

B. K. Ghosh is with the Department of Systems Science and Mathematics, Washington University, St. Louis, MO 63130 USA.

J. Rosenthal is with the Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556 USA.

IEEE Log Number 9406983.

On the other hand, if 4 points are observed the output vector in (1.1) is observed up to a homogeneous line. Various other cases can be demonstrated likewise. Note in particular that by observing the vector $(\eta_1, \eta_2, \eta_3, \eta_4)$ for 1 or 2 points on the image plane, the output vector is observed up to respectively a seven- or five-dimensional plane in \mathbb{R}^9 . Such an observation does not shed any new information on the vector (p, q, s, t) . In practice if the vector (p, q, s, t) is recovered only up to a d -dimensional plane where $d > 1$ one would typically use multiple cameras to determine the exact value of (p, q, s, t) .

II. PROBLEM FORMULATION AND MAIN RESULT

In order to introduce the main result considered in this paper let $P_0 \subset \mathbb{K}^n$ be a d -dimensional affine subspace not necessarily passing through the origin. In this paper we shall use the expression ' d -dimensional affine subspace' to mean a ' d -dimensional plane'. We say that the dynamical system (1.1) observes P_0 if for any $0 \leq t_1 < t$ it is possible to calculate P_0 from the observation of the moving plane $C(P(t)) \triangleq C e^{-\lambda t} P_0$ in \mathbb{K}^n , $t_1 < t \leq t$. Our main theorem is described as follows.

Theorem 2 (Main Theorem) System (1.1) observes any d -dimensional affine subspace P_0 in \mathbb{K}^n if for any set of eigenvalues $\lambda_1, \dots, \lambda_d$ of A one has

$$\text{rank} \begin{bmatrix} (A - \lambda_1 I) & \cdots & (A - \lambda_d I) \\ C \end{bmatrix} = n \quad (2.1)$$

Moreover this condition is also necessary if $d = 0$ or if all eigenvalues of the matrix A are in \mathbb{K} .

Remark 3 Note that over the complex numbers \mathbb{C} condition (2.1) is necessary and sufficient. Moreover if $d = 0$ Theorem 2 reduces to Theorem 1. Finally if $d = 1$ this result implies the one given in [6] due to Wang, Martin, Dayawansa, and Ghosh.

The following two examples explain the ingredients of our result.

Example 4 Consider the real system

$$\dot{x} = Ax = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x, \quad y = Cx = (1 \ 0)x, \quad x \in \mathbb{R}^2 \quad (2.2)$$

Because the eigenvalues of A are real, condition (2.1) in Theorem 2 is necessary and sufficient. In particular if $p_0 \in \mathbb{R}^2$ is a point it can be observed from $y(t) = C e^{-\lambda t} p_0$ because

$$\text{rank} \begin{bmatrix} -\lambda & 1 \\ 0 & -\lambda \\ 1 & 0 \end{bmatrix} = 2 \quad (2.3)$$

for all $\lambda \in \mathbb{R}$ including the case when λ is an eigenvalue of A . However if $l_0 \subset \mathbb{R}^2$ is a line, it cannot be observed from $l(t) = C e^{-\lambda t} l_0$ because for every pair of lines l_0 and l_1 in \mathbb{R}^2 and for all but a finite number of time instants t we have

$$\{\xi \mid \xi = C e^{-\lambda t} \delta \in l_0\} = \{\xi \mid \xi = C e^{-\lambda t} \delta \in l_1\} \quad (2.4)$$

Thus the lines l_0 and l_1 are both mapped to the entire real axis and therefore they cannot be observed. We also note that

$$\text{rank} \begin{bmatrix} \lambda_0 \lambda_1 & -\lambda_0 - \lambda_1 \\ 0 & \lambda_0 \lambda_1 \\ 1 & 0 \end{bmatrix} \neq 2 \quad \text{for every pair of eigenvalues } \lambda_0, \lambda_1 \text{ of } A \quad (2.5)$$

In fact for $\lambda_0 = \lambda_1 = 0$ rank drops to 1.

Remark 5 Note that the equality of the two sets l_0 and l_1 for all but possibly a finite number of time instants t does not correctly conclude from this that in principle observations can be ascertained on the basis of these finitely many t . However we would still like to say that the lines l_0 and l_1 are unobservable on the basis of any arbitrary time interval.

Example 6 Consider the fourth order system

$$\dot{x} = Ax = \begin{pmatrix} -81 & -56 & 57 & 11 \\ 146 & 102 & -106 & 20 \\ 62 & 13 & -46 & 9 \\ 203 & 138 & -149 & 31 \end{pmatrix} x, \quad y = Cx = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} x, \quad x \in \mathbb{R}^4 \quad (2.6)$$

A direct computation shows that the pair (A, C) is observable and the matrix A has eigenvalues $0, 1, 2, 3$. Since for any 2 eigenvalues λ_0, λ_1 the nullspace of $(A - \lambda_0 I)(A - \lambda_1 I)$ is equal to the sum of the eigenspaces $\text{Ker}(A - \lambda_0 I)$ and $\text{Ker}(A - \lambda_1 I)$ and none of those sums contains the vector $(0, 0, 0, 1)^T \in \text{Ker}(C)$ it follows from Theorem 2 that if $l_0 \subset \mathbb{R}^4$ is a line in \mathbb{R}^4 it can be observed from $l(t) = C e^{-\lambda t} l_0$ which is a motion of lines in \mathbb{R}^4 . On the other hand one immediately verifies that

$$l(A - I)(A - 2I) = \begin{pmatrix} -3 & -1 & 3 & 0 \\ 8 & 8 & -8 & 0 \\ -1 & -1 & 1 & 0 \\ -23 & -23 & 23 & 0 \end{pmatrix}$$

It therefore follows that certain two planes $P_0 \subset \mathbb{R}^4$ cannot be observed from $C e^{-\lambda t} P_0 \subset \mathbb{R}^4$. Specifically consider the vectors $v_0 = (-12, 20, 8, 28)^T$, $v_1 = (35, -60, -25, -55)^T$ and $v_2 = (-23, 40, 17, 58)^T$. One immediately verifies that v_0, v_1, v_2 are eigenvectors corresponding to the eigenvalues $0, 1$ and 2 . Also note that $v_0 + v_1 + v_2 = (0, 0, 0, 1)^T$. Let P be the three dimensional subspace in \mathbb{R}^4 spanned by the vectors v_0, v_1 and v_2 . It can be verified that for all but a finite set of values of t $C e^{-\lambda t} P$ is a two dimensional plane in \mathbb{R}^4 . To see this note that $C e^{-\lambda t} v_j = C e^{-\lambda_j t} v_j$ for $j = 0, 1, 2$ where λ_j is the eigenvalue corresponding to the eigenvector v_j . Since $C v_0, C v_1, C v_2$ are linearly dependent it follows that $C e^{-\lambda t} v_0, C e^{-\lambda t} v_1, C e^{-\lambda t} v_2$ are linearly dependent as well. Thus for any $0 \leq t_1 < t$ and for almost every pair of two dimensional planes Q_1 and Q_2 such that $Q_1 \neq Q_2$ and $Q_1 \subset P, Q_2 \subset P$ we have

$$C e^{-\lambda t} Q_1 = C e^{-\lambda t} Q_2$$

for $t_1 \leq t \leq t$. Hence the planes Q_1 and Q_2 cannot be observed.

III. AN ASSOCIATED DYNAMICAL SYSTEM

The proof of Theorem 2 will be broken down in a sequence of lemmas. The proof is mainly based on a careful study of a dynamical system defined on the \mathbb{K} vector space $\wedge^k \mathbb{K}^n$ the k -fold wedge product of \mathbb{K}^n (see [2] for a reference). This system has also been used in [6] to derive the results there.

First recall that $\wedge^k \mathbb{K}^n$ is linearly generated by the vectors

$$\{e_{i_1} \wedge \cdots \wedge e_{i_k} \mid i_1 < \cdots < i_k\}$$

Addition in $\wedge^k \mathbb{K}^n$ is multilinear and alternating in the components. If $\{e_{i_1}, \dots, e_{i_k}\}$ is a basis of \mathbb{K}^n then it follows from the multilinearity and the alternating property of the wedge product that

$$\mathcal{B} = \{e_{i_1} \wedge \cdots \wedge e_{i_k} \mid 1 \leq i_1 < \cdots < i_k \leq n\}$$

is a basis of $\bigwedge^k \mathbb{K}^n$. In particular one has $\dim \bigwedge^k \mathbb{K}^n = \binom{n}{k}$. If a vector $v \in \bigwedge^k \mathbb{K}^n$ has a representation $v = x_1 \wedge \cdots \wedge x_k$ for some particular vectors $x_i \in \mathbb{K}^n$, $i = 1, \dots, k$, one says that v is a *decomposable* vector. The coordinates of a decomposable vector with respect to the canonical basis \mathcal{B} are sometimes called the Plücker coordinates of v .

Next define linear maps

$$\begin{aligned} \hat{A}: \bigwedge^k \mathbb{K}^n &\rightarrow \bigwedge^k \mathbb{K}^n \\ x_1 \wedge \cdots \wedge x_k &\mapsto \sum_{i=1}^k x_1 \wedge \cdots \wedge x_{i-1} \wedge Ax_i \wedge x_{i+1} \wedge \cdots \wedge x_k \end{aligned} \quad (3.1)$$

and

$$\begin{aligned} C': \bigwedge^k \mathbb{K}^n &\rightarrow \bigwedge^k \mathbb{K}^q \\ x_1 \wedge \cdots \wedge x_k &\mapsto C'x_1 \wedge \cdots \wedge C'x_k. \end{aligned} \quad (3.2)$$

\hat{A} and C' induce the dynamical system

$$\dot{X} = \hat{A}X, \quad \dot{Y} = C'X. \quad (3.3)$$

The state space of (3.3) is the vector space $\bigwedge^k \mathbb{K}^n$ and the output space is the vector space $\bigwedge^k \mathbb{K}^q$. We would like to remark that if the trajectories $C'x_1(t), \dots, C'x_k(t)$ are solutions of the system (1.1) then $\hat{C}(x_1(t) \wedge \cdots \wedge x_k(t))$ is a solution of system (3.3). It is our goal to show that, provided the eigenvalues of A are in \mathbb{K} , (2.1) is equivalent to a particular notion of observability of the system (3.3) and that this condition is also necessary and sufficient for the observability of P_0 under the output function $C'e^{-t}P_0$. The following lemmas prepare for this result.

Lemma 7: The (unique) solution of the initial value problem

$$\begin{aligned} \frac{d}{dt}(x_1(t) \wedge \cdots \wedge x_k(t)) &= \hat{A}(x_1(t) \wedge \cdots \wedge x_k(t)) \\ x_1(0) \wedge \cdots \wedge x_k(0) &= v_1 \wedge \cdots \wedge v_k \end{aligned}$$

is given through

$$x_1(t) \wedge \cdots \wedge x_k(t) = e^{t\hat{A}} v_1 \wedge \cdots \wedge e^{t\hat{A}} v_k. \quad (3.4)$$

Proof: Differentiate (3.4) and recall the definition of \hat{A} . Q.E.D.

Lemma 8: Let $x_1, \dots, x_k \in \mathbb{K}^n$ be vectors and $c_1, c_2, \dots, c_k \in \mathbb{K}$ be scalars. Let $v \triangleq c_1 + c_2 + \cdots + c_k$. Then it follows that

$$\begin{aligned} (\hat{A} - vI)(x_1 \wedge x_2 \wedge \cdots \wedge x_k) &= \sum_{i=1}^k x_1 \wedge \cdots \wedge x_{i-1} \wedge (\hat{A} - c_i I)x_i \wedge x_{i+1} \wedge \cdots \wedge x_k. \end{aligned} \quad (3.5)$$

Proof. Note that

$$\begin{aligned} &(\hat{A} - vI)(x_1 \wedge x_2 \wedge \cdots \wedge x_k) \\ &= (Ax_1 \wedge x_2 \wedge \cdots \wedge x_k) \\ &\quad - c_1(x_1 \wedge x_2 \wedge \cdots \wedge x_k) \\ &\quad + (x_1 \wedge Ax_2 \wedge \cdots \wedge x_k) \\ &\quad - c_2(x_1 \wedge x_2 \wedge \cdots \wedge x_k) \\ &\quad + \cdots \\ &\quad - c_k(x_1 \wedge x_2 \wedge \cdots \wedge x_k) \\ &= \sum_{i=1}^k x_1 \wedge \cdots \wedge x_{i-1} \wedge (\hat{A} - c_i I)x_i \wedge x_{i+1} \wedge \cdots \wedge x_k. \end{aligned} \quad \text{Q.E.D.}$$

Lemma 9: Let $\{x_1, \dots, x_n\} \subset \mathbb{K}^n$ be a \mathbb{K} -basis of generalized eigenvectors of the matrix A having corresponding eigenvalues $\{\lambda_1, \dots, \lambda_n\}$ (possibly repeated) then

$$\{x_{i_1} \wedge \cdots \wedge x_{i_k} \mid 1 \leq i_1 < \cdots < i_k \leq n\} \quad (3.6)$$

is a \mathbb{K} -basis of generalized eigenvectors of the matrix \hat{A} with corresponding eigenvalues $\lambda_{i_1} + \cdots + \lambda_{i_k}$.

Proof: Clearly the set of vectors (3.6) are linearly independent and therefore form a basis. Assume that the vectors x_{i_1}, \dots, x_{i_k} have a nilpotency index m_{i_1}, \dots, m_{i_k} , i.e.,

$$(A - \lambda_{i_k} I)^{m_{i_k}-1} x_{i_k} \neq 0, \quad (A - \lambda_{i_k} I)^{m_{i_k}} x_{i_k} = 0. \quad (3.7)$$

In particular, if $m_{i_k} = 1$ it follows that x_{i_k} is an eigenvector with λ_{i_k} being the corresponding eigenvalue. Let us define

$$q = m_{i_1} + \cdots + m_{i_k} + 1 - k \quad (3.8)$$

it is trivial to verify using Lemma 8 that

$$(\hat{A} - (\lambda_{i_1} + \cdots + \lambda_{i_k})I)^q x_{i_1} \wedge \cdots \wedge x_{i_k} = 0. \quad (3.9)$$

Q.E.D.

Lemma 10: Let N be a nilpotent operator acting on \mathbb{K}^q . For every vector $v \in \mathbb{K}^q$ there is a unique $u \in \mathbb{K}^q$ such that

$$v = u + Nu + \cdots + N^{q-1}u.$$

Moreover if m is the nilpotency index of v then $\{u, \dots, N^{m-1}u\}$ are linearly independent.

Proof: The unique vector u is given through $u := (I - N)v$. The linear independence is clear. Q.E.D.

Before we state the next result we note the following.

Remark 11: Note that not every vector in the vector space $\bigwedge^k \mathbb{K}^n$ is of the form $x_1 \wedge \cdots \wedge x_k$ and those that are, would be known as decomposable vectors.

The next result establishes the crucial relation between the observability of the pair (\hat{A}, C') and the condition (2.1).

Proposition 12: Assume that the eigenvalues of the matrix A are in \mathbb{K} . Then the following conditions are equivalent:

1) There are eigenvalues $\lambda_{i_1}, \dots, \lambda_{i_k}$ of A and a nonzero vector $v \in \mathbb{K}^n$ such that

$$\left((A - \lambda_{i_1} I) \cdots (A - \lambda_{i_k} I) \right)_{C'} v = 0. \quad (3.10)$$

2) There is a $\lambda \in \mathbb{K}$ and a decomposable vector $z_1 \wedge \cdots \wedge z_k \in \bigwedge^k \mathbb{K}^n$ such that

$$\left(\frac{A - \lambda I}{C'} \right) z_1 \wedge \cdots \wedge z_k = 0. \quad (3.11)$$

3) The dynamical system (3.3) has a decomposable vector $\alpha_1 \wedge \cdots \wedge \alpha_k \in \bigwedge^k \mathbb{K}^n$ in its unobservable subspace.

Proof: 1) \rightarrow 2): Let $\lambda_1, \dots, \lambda_k$ be the eigenvalues of A and let

$$\mathbb{K}^n = \bigoplus_{i=1}^s V_{\lambda_i} \quad (3.12)$$

be the decomposition into generalized eigenspaces. This decomposition induces a decomposition

$$v = v_1 + \cdots + v_s.$$

Let $\Lambda = \{\lambda_1, \dots, \lambda_p\}$ be the eigenvalues appearing in the product

$$P := (A - \lambda_{i_1} I) \cdots (A - \lambda_{i_k} I)$$

and denote by $m(\lambda_1), \dots, m(\lambda_p)$ their multiplicity, i.e., we have

$$P = (A - \lambda_1 I)^{m(\lambda_1)} \cdots (A - \lambda_p I)^{m(\lambda_p)}.$$

From the A invariance of the generalized eigenspaces V_{λ_i} it follows that $v_h = 0$ if $h \notin \Lambda$. Moreover if $h \in \Lambda$ then v_h has nilpotency

index at most $m(h)$. In the following we restrict ourselves to the case when v_h has nilpotency index $m(h)$. The (easier) other cases are similar. By Lemma 10 we have an expansion

$$v = \sum_{j_1=0}^{m(l_1)-1} (A - l_1 I)^{j_1} u_{l_1} + \cdots + \sum_{j_p=0}^{m(l_p)-1} (A - l_p I)^{j_p} u_{l_p}. \quad (3.13)$$

In this summation there are $m(l_1) + \cdots + m(l_p) = k$ summands which we like to denote by j_1, \dots, j_k . By Lemma 10 those vectors are linearly independent and from Lemma 9 it follows that $j := j_1 \wedge \cdots \wedge j_k$ is an eigenvector of \hat{A} with corresponding eigenvalue $\lambda := \lambda_{l_1} + \cdots + \lambda_{l_k}$. Finally, from (3.10) it follows that $\{Cj_1, \dots, Cj_k\}$ is a linearly dependent set. It follows

$$\left(\hat{A} - (\lambda_{l_1} + \cdots + \lambda_{l_k})I \right) j_1 \wedge \cdots \wedge j_k = 0. \quad (3.14)$$

2) \rightarrow 3): The vector $j_1 \wedge \cdots \wedge j_k$ is necessarily an eigenvector of \hat{A} and therefore in the unobservable subspace U of the system (3.3).

3) \rightarrow 1): The fact that condition 3) implies condition 1) is nontrivial. Our proof follows mainly ideas already developed in [6] and in principle it should be possible to generalize the proof given in [6]. This however amounts to a large case by case search. In order to avoid those tedious arguments we will deviate at a crucial point from this program.

The proof is structured as follows. Consider the decomposable vector $\alpha := \alpha_1 \wedge \cdots \wedge \alpha_k$ in the unobservable subspace U of the system (3.3) whose existence we assume. Using the fact that U is \hat{A} invariant we will construct a polynomial $f(x) \in \mathbb{K}[x]$ which has the property that $f(\hat{A})\alpha$ is a decomposable eigenvector of \hat{A} . $f(\hat{A})\alpha$ is then necessarily in the unobservable subspace U and this implies 2) and from there we will imply 1). The details are now described as follows.

Consider the set of eigenvalues $\{\lambda_1, \dots, \lambda_k\} \subset \mathbb{K}$ of \hat{A} and arrange the order such that

$$\operatorname{Re} \lambda_i < \operatorname{Re} \lambda_{i+1}$$

or

$$\operatorname{Re} \lambda_i = \operatorname{Re} \lambda_{i+1}$$

and

$$\operatorname{Im} \lambda_i < \operatorname{Im} \lambda_{i+1}.$$

Let us choose a set of generalized eigenvectors $\{x_1, \dots, x_n\}$ of \hat{A} and consider the decomposition of \mathbb{K}^n into generalized eigenspaces given through $\mathbb{K}^n = \bigoplus_{i=1}^k V_{\lambda_i}$. Arrange the order of $\{x_1, \dots, x_n\}$ in such a way that $\{x_1, \dots, x_{l_1}\}$ forms a basis of V_{λ_1} , $\{x_{l_1+1}, \dots, x_{l_1+l_2}\}$ forms a basis of V_{λ_2} and so on.

Let $\alpha := \alpha_1 \wedge \cdots \wedge \alpha_k$ be a decomposable vector in the unobservable subspace U . Expand $\alpha_j = \sum_{i=1}^n b_{ji} x_i$, $j = 1, \dots, k$, in terms of this basis. In this way we associate to α a coefficient matrix $B = b_{ij}$ whose entries are unique up to premultiplication by an element of the special linear group $SL_k := \{T \in \operatorname{Mat}_{k \times k} | \det(T) = 1\}$. Without loss of generality we can therefore assume that the matrix B is in echelon form.

Consider now the decomposition of $\bigwedge^k \mathbb{K}^n$ into generalized eigenspaces.

$$\bigwedge^k \mathbb{K}^n = \bigoplus_{\lambda = \lambda_{i_1} + \cdots + \lambda_{i_k}} W_{\lambda}. \quad (3.15)$$

If w_{λ} denotes the component of $\alpha_1 \wedge \cdots \wedge \alpha_k$ in W_{λ} then

$$w_{\lambda} = \sum v_{i_1} \wedge \cdots \wedge v_{i_k}, \quad (3.16)$$

where v_{i_j} is the component of α_j in $V_{\lambda_{i_j}}$ and where the is taken over all indexes (i_1, \dots, i_k) having the property $\lambda_{i_1} + \cdots + \lambda_{i_k} = \lambda$.

From the fact that the matrix B is assumed to be in echelon and from the assumption that $\lambda_1, \dots, \lambda_k$ are ordered, it follows that there is one eigenvalue μ such that the component w_{μ} of $\alpha_1 \wedge \cdots \wedge \alpha_k$ in W_{μ} is nonzero and decomposable, i.e.,

$$w_{\mu} = v_{i_1} \wedge \cdots \wedge v_{i_k}. \quad (3.17)$$

Indeed, v_{i_j} can be chosen in the following way. For $j = 1, \dots, k$ consider the decomposition

$$\alpha_j = \sum_{i_j=1}^{n_j} v_{i_j} \quad (3.18)$$

induced by the eigenspace decomposition (3.12). Then choose i_j as the first index with the property, that $v_{i_j} \neq 0$. By definition $v_{i_1} \wedge \cdots \wedge v_{i_k}$ is nonzero, decomposable, and it represents the component of $\alpha_1 \wedge \cdots \wedge \alpha_k$ in W_{μ} .

Let m be the order of w_{μ} . It is our goal to calculate the eigenvector $(\hat{A} - \mu I)^{m-1} w_{\mu}$ and to show that this vector is decomposable as well. For this consider the initial value problem

$$\begin{aligned} \frac{d}{dt}(x_1(t) \wedge \cdots \wedge x_k(t)) &= (A - \mu I)(x_1(t) \wedge \cdots \wedge x_k(t)) \\ x_1(0) \wedge \cdots \wedge x_k(0) &= v_{i_1} \wedge \cdots \wedge v_{i_k}. \end{aligned}$$

Using Lemma 7 and Lemma 8 one verifies that the solution is given through

$$e^{(A - \mu I)t} w_{\mu} = e^{(\lambda_{i_1} - \mu)t} v_{i_1} \wedge \cdots \wedge e^{(\lambda_{i_k} - \mu)t} v_{i_k}. \quad (3.19)$$

Because $v_{i_j} \in V_{\lambda_{i_j}}$, $j = 1, \dots, k$, we have a polynomial expansion

$$e^{(\lambda_{i_j} - \mu)t} v_{i_j} = \sum_{h_j=0}^{m(j)-1} (A - \lambda_{i_j} I)^{h_j} t^{h_j} v_{i_j}, \quad (3.20)$$

where $m(j)$ is the nilpotency index of v_{i_j} .

Expanding each v_{i_j} in terms of the standard basis $\{x_1, \dots, x_n\} \subset \mathbb{K}^n$ we get an expansion

$$e^{(A - \mu I)t} w_{\mu} = \sum_{1 \leq i_1 < \cdots < i_k \leq n} f_{(i_1, \dots, i_k)}(t) x_{i_1} \wedge \cdots \wedge x_{i_k}. \quad (3.21)$$

In this summation $f_{(i_1, \dots, i_k)}(t)$ are the Plücker coordinates of the vector $e^{(A - \mu I)t} w_{\mu}$ and we will abbreviate them by $f_i(t)$. By (3.20) it follows that $f_i(t)$ are all polynomials of degree at most $\sum_{j=1}^k m(j) - k$.

If fact we can say more. Differentiating both sides in (3.19) $m-1$ times and substituting $t = 0$ results in the eigenvector $(A - \mu I)^{m-1} w_{\mu}$ on the left side of the equality sign. On the right side this operation results in

$$(m-1)! \sum_{1 \leq i_1 < \cdots < i_k \leq n} g_i x_{i_1} \wedge \cdots \wedge x_{i_k},$$

where g_i is the coefficient of the monom t^{m-1} in the polynomial $f_i(t)$. By definition we have $(A - \mu I)^{m-1} w_{\mu} \neq 0$ and $(A - \mu I)^m w_{\mu} = 0$. We conclude that each polynomial $f_i(t)$ has degree at most $m-1$ and some coefficients g_i are nonzero. In addition note that the vector $e^{(A - \mu I)t} w_{\mu}$ is a decomposable vector at all time $t \geq 0$ and the Plücker relations (compare to [5, Section 3])

$$(QR) \quad \sum_{n-k+1}^{p+1} (-1)^k \cdot f_{(i_1, \dots, i_{p-k+1}, j_k)}(t) \cdot f_{(j_1, \dots, j_{k-1}, i_{p-k+1})}(t) = 0 \quad (3.22)$$

have therefore to be satisfied for all $t \geq 0$ as well. By doing the same argument as in [5, Theorem 3.6 and Example 3.7] we conclude

that the Plücker coordinates are also satisfied for the coordinates g_i . But this means that $(A - \mu I)^{m-1} w_\mu$ is a decomposable eigenvector which we denote by

$$j_1 \wedge \cdots \wedge j_k.$$

Consider once more the eigenspace decomposition

$$\alpha_1 \wedge \cdots \wedge \alpha_k = \sum_{\lambda} w_\lambda$$

as induced by the decomposition (3.15). Let $m(\lambda)$ be the nilpotency index of w_λ and define the operator

$$f(\dot{A}) := (A - \mu I)^{m-1} \prod_{\lambda \neq \mu} (A - \lambda I)^{m(\lambda)}. \quad (3.23)$$

A direct calculation shows that

$$f(\dot{A})\alpha_1 \wedge \cdots \wedge \alpha_k = \prod_{\lambda \neq \mu} (\mu - \lambda)^{m(\lambda)} j_1 \wedge \cdots \wedge j_k. \quad (3.24)$$

We conclude that the decomposable eigenvector $j_1 \wedge \cdots \wedge j_k$ is in the unobservable subspace U of the system (3.3) and this implies 2).

Actually we have shown more: $j_i \in V_i$, and if the coefficient r_j is repeated m times in the set $\{r_1, \dots, r_k\}$ then j_i has nilpotency index at most m . But this means that

$$(A - \lambda_{j_1} I) \cdots (A - \lambda_{j_k} I) j_i = 0$$

for $j = 1, \dots, k$.

By linear dependence of the set $\{C' j_1, \dots, C' j_k\}$ there exist scalars c_1, \dots, c_k not all zero such that

$$c_1 C' j_1 + \cdots + c_k C' j_k = 0.$$

But then

$$w \triangleq c_1 j_1 + \cdots + c_k j_k$$

has all required properties for 1).

Q.E.D.

A direct consequence is the following Lemma whose proof is clear.

Lemma 13: If for any set of eigenvalues $\lambda_1, \dots, \lambda_k$ of A one has

$$\text{rank} \begin{bmatrix} (A - \lambda_1 I) \cdots (A - \lambda_k I) \\ C' \end{bmatrix} = n \quad (3.25)$$

then there is no real decomposable vector in the unobservable subspace U of the system (3.3).

Remark 14: In general the converse is not true as it is demonstrated through an example in [6].

IV. PROOF OF THE MAIN THEOREM

Proof: We first show the sufficiency of the criterion (2.1). Let $P_0, Q_0 \subseteq \mathbb{K}^n$ be two d -dimensional planes with $P_0 \neq Q_0$. Let $q_0 \in Q_0$ be a point having the property that $q_0 \notin P_0$. Let $\{x_0, \dots, x_d\} \subseteq P_0$ be a set of points chosen in such a way that the decomposable vector

$$w \triangleq (q_0 - x_0) \wedge (x_1 - x_0) \wedge \cdots \wedge (x_d - x_0)$$

is nonzero. If the rank condition (2.1) holds, it follows from Proposition 12 and Lemma 13 that there is no decomposable vector in the unobservable subspace U of "the augmented system" (3.3). It therefore follows that

$$C'w(t) = (C'e^{-At}q_0 - C'e^{-At}x_0) \wedge (C'e^{-At}x_1 - C'e^{-At}x_0) \cdots \wedge (C'e^{-At}x_d - C'e^{-At}x_0)$$

is nonzero for all time t with the exception of a set of measure zero. But then we have that $C'e^{-At}q_0 \notin C'e^{-At}P_0$ for almost all time t . In other words $C'e^{-At}Q_0 \neq C'e^{-At}P_0$.

In order to prove the necessity part assume that all eigenvalues of A are in \mathbb{K} . Assume that there is a set of eigenvalues $\lambda_0, \dots, \lambda_d$ of A such that the rank condition (2.1) is not satisfied. Furthermore assume that for any set of eigenvalues μ_0, \dots, μ_{d-1} of A

$$\text{rank} \begin{bmatrix} (A - \mu_0 I) \cdots (A - \mu_{d-1} I) \\ C' \end{bmatrix} = n. \quad (4.1)$$

If this (technical) condition is not satisfied we will be able to show at the end of the proof that $(d-1)$ -dimensional subspaces cannot be observed in general.

By Proposition 12 there exists a nonzero, decomposable vector $x_0 \wedge x_1 \wedge \cdots \wedge x_d$ in the unobservable subspace U of "the augmented system" (3.3) (assuming $k = d+1$). Define $V := \text{span}\{x_0, \dots, x_d\}$ and let $P_0, Q_0 \subset V$ be two d -dimensional subspaces satisfying $P_0 \neq Q_0$. By the assumption it follows that

$$\text{span}\{C'e^{-At}x_0, \dots, C'e^{-At}x_d\}$$

is a d -dimensional subspace for t almost everywhere. But this means that the two different subspaces P_0 and Q_0 in \mathbb{K}^n produce the same moving plane $C'e^{-At}P_0 = C'e^{-At}Q_0$ in \mathbb{K}^n for all time t with the possible exception of a set of measure zero.

Assume now that (4.1) is not satisfied and let \hat{d} be the largest integer having the property that

$$\text{rank} \begin{bmatrix} (A - \lambda_0 I) \cdots (A - \lambda_{\hat{d}} I) \\ C' \end{bmatrix} < n \quad (4.2)$$

for some eigenvalues $\lambda_0, \dots, \lambda_{\hat{d}}$ but

$$\text{rank} \begin{bmatrix} (A - \mu_0 I) \cdots (A - \mu_{\hat{d}-1} I) \\ C' \end{bmatrix} = n \quad (4.3)$$

for all eigenvalues $\mu_0, \dots, \mu_{\hat{d}-1}$ of A . Using the same argument as before one shows the existence of two subspaces \tilde{P}_0 and \tilde{Q}_0 of dimension \hat{d} which cannot be distinguished in the observation. This completes the proof. Q.E.D.

ACKNOWLEDGMENT

The first author would like to thank Prof. C. Martin for many interesting advice on the perspective observability problem. Comments made by three anonymous referees on an earlier version of this paper is gratefully acknowledged. Finally we would like to thank A. A. Stoorvogel for making numerous suggestions and comments to improve the presentation of the paper.

REFERENCES

- [1] B. K. Ghosh, M. Jankovic, and Y. T. Wu, "Perspective problems in system theory and its application to machine vision," *J. Math. Syst. Estimation Contr.*, vol. 4, no. 1, pp. 3-38, 1994.
- [2] V. I. Arnold, *Mathematical Methods of Classical Mechanics*. New York: Springer-Verlag, 1978.
- [3] M. L. J. Hautus, "Controllability and observability condition of linear autonomous systems," *Ned. Akad. Wetenschappen, Proc. Ser. A*, vol. 72, pp. 443-448, 1969.
- [4] B. K. P. Horn, *Robot Vision*. Cambridge, MA: MIT Press, 1986.
- [5] J. Rosenthal, "On dynamic feedback compensation and compactification of systems," *SIAM J. Contr. Optim.*, vol. 32, no. 1, pp. 279-296, 1994.
- [6] W. P. Dayawansa, B. K. Ghosh, C. Martin, and X. Wang, "A necessary and sufficient condition for the perspective observability problem," *Syst. Contr. Lett.*, to be published.

A Note on Robust Pole Placement

Marek K. Solak and Albert C. Peng

Abstract— Pole placement algorithm in single input–single output (SISO) systems is discussed with respect to the corresponding, real stability radius of the resulting closed loop polynomial

I. INTRODUCTION

Various robust stability problems are discussed in terms of stability radius [1], [2], [9]–[14]. This approach is attractive because only one variable (stability radius) describes stability margin in a parameter space. Numerous methods exist to calculate stability radius (structured and unstructured) [1], [2], [9]–[14] and in many cases these methods allow to determine destabilizing direction of perturbation. Real stability radius appears also as a practical tool in designing robust stabilizing controllers for various control systems [1], [2].

This note provides a useful suggestion how pole placement of a closed loop system affects robust stability measured by its real stability radius. It is also interesting to notice when root localization of a given Hurwitz polynomial is known then complicated numerical minimization of its distance d_∞ from the set of polynomials with roots on the imaginary axis can be avoided in many cases. d_∞ may have many spikes and local minima over $[0, \infty)$.

Consider a nominal Hurwitz polynomial

$$p_{i-1}(s) = a_0^{i-1}s^n + a_1^{i-1}s^{n-1} + \dots + a_{n-1}^{i-1}s + a_n^{i-1} = p_{i-1}'(s), \quad S(s) \quad (1)$$

$$S(s) = [s^n \quad s^{n-1} \quad \dots \quad s \quad 1] \quad (2)$$

perturbed by a polynomial

$$h(s) = \delta_0 s^n + \delta_1 s^{n-1} + \dots + \delta_{n-1} s + \delta_n = \delta' S(s) \quad (3)$$

The real unstructured stability radius for $p_{i-1}(s)/p_{i-1}'(s)$ can be calculated via various methods. All of these methods are based on the paper of Fam and Meditch [16]. Define

$$p_{i-1}(s) = p_{i-1}'(s) + \kappa p_{i-1}''(s) \quad (4)$$

Theorem 1 [1], [9]–[14]

$$r_{i-1} = r_{i-1}(p_{i-1}(s)) = \min(|a_0^{i-1}|, |a_n^{i-1}|, d_\infty) \quad (5)$$

$$d_\infty = \inf_{\omega} \left\{ \frac{[p_{i-1}(j\omega)]^2}{1 + \omega^{2n} + \dots + \omega^{2(n-1)}} + \frac{[p_{i-1}(j\omega)]^2}{1 + \omega^{2n} + \dots + \omega^{2(n-1)}} \right\} \quad n = 2p \quad (6a)$$

$$d_\infty = \inf_{\omega} \left\{ \frac{[p_{i-1}(j\omega)]^2}{1 + \omega^{2n} + \dots + \omega^{2(n-1)}} + \frac{[p_{i-1}(j\omega)]^2}{1 + \omega^{2n} + \dots + \omega^{2(n-1)}} \right\} \quad n = 2p + 1 \quad (6b)$$

where $|a_0^{i-1}|$ denotes the distance of $p_{i-1}(s)$ from a zero at infinity, $|a_n^{i-1}|$ denotes the distance from a zero at the origin, and d_∞ denotes

Manuscript received December 8, 1993; revised May 15, 1994. The work of M. K. Solak was supported by the Foundation for Research Development of the Republic of South Africa under Grant 2015440.

M. K. Solak is with the Department of Fundamental Research in Electrical Engineering, Polish Academy of Sciences and Ministry of Industry, Institute of Electrical Engineering, 04-703 Warszawa, ul. Polzaryskiego 28, Poland.

A. C. Peng is with the Department of Industrial and Engineering Technology, Central Michigan University, Mt. Pleasant, MI 48859, USA.

IEEE Log Number 9406984.

the distance from the set of polynomials with roots on the imaginary axis at $\pm j\omega$.

In many pole placement algorithms we are able to calculate [6], localization of $p_{i-1}(s)$ via (state space) feedback, closed loop polynomial

$$p_i(s) = p_{i-1}(s) - f(s)$$

is generated [1], [3]. Some other feedback controllers (cf. Section II) require solution of a diophantine polynomial equation

$$p(s)a(s) + q(s)b(s) = n_{i-1}(s)$$

where $n_{i-1}(s)$ is a polynomial with arbitrary location of roots within the left complex half plane. At this stage an open question remains of how to select either $f(s)$ or $n_{i-1}(s)$ to assure maximum value of the corresponding stability radius. On the other hand, many practical problems impose specific regions (cf. [7]) in the left complex half plane where roots of $p_i(s)$ or $n_{i-1}(s)$ shall be located.

This note discusses a connection between real unstructured stability radius of a given Hurwitz polynomial and its root localization.

II. ROBUST POLE PLACEMENT

Presentation of the main result of this note will require additional definitions. Let $S^-(p)$ denote the set of all Hurwitz polynomials with roots outside interior of the unit circle (single or multiple roots on the circumference of the unit circle are allowed). Additionally let $\mathcal{HP}^-(p)$ denote the set of all Hurwitz polynomials with roots in the region bounded by the hyperbola (including roots on the hyperbola)

$$\frac{\alpha^2}{\left(\frac{1}{\sqrt{\alpha}}\right)^2} - \frac{\beta^2}{\left(\frac{1}{\sqrt{\alpha}}\right)^2} = 1 \quad (7)$$

where α denotes real part of the root and β denotes complex part of the root.

The main result of this note is formulated in the following theorem. The proof is a little complicated and is carried out in the next section.

Theorem 2 If

$$p_{i-1}(s) \in S^-(p) \cap \mathcal{HP}^-(p)$$

then

$$r_{i-1}(p_{i-1}(s)) = |a_0^{i-1}| \quad (8)$$

Remark 1

The condition $p_{i-1}(s) \in S^-(p)$ can be easily verified. Assuming that $p_{i-1}(s)$ has no roots at infinity we can easily see that $p_{i-1}(s) \in S^-(p)$ iff

$$q(s) = s p_{i-1}(1/s) \in S^-(p)$$

where $S^-(p)$ denotes the set of all Schur polynomials (polynomials with all roots inside the unit circle). There is a fair number of algebraic procedures to check if $q(s) \in S^-(p)$ (cf. [4], [15]). Applying the method in [15] we are also able to check if $p_{i-1}(s) \in \mathcal{HP}^-(p)$.

It is important to notice that stabilization algorithms are mainly based on an *a priori* localization of roots of a nominal closed loop system characteristic polynomial within the left complex half plane and do not require application of G stability polynomial tests (cf. [15]).

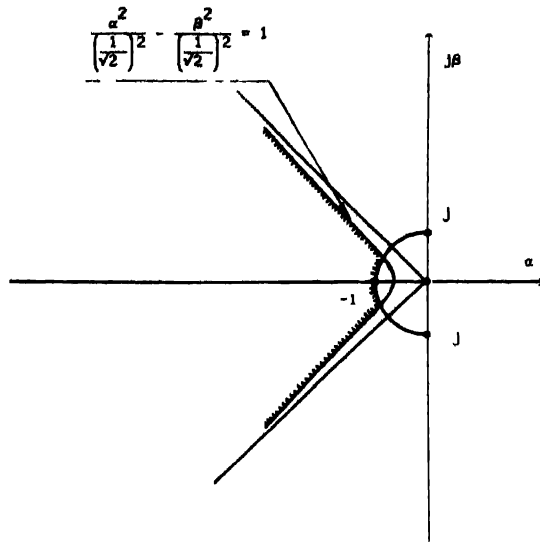


Fig. 1 Root localization of polynomials destabilized through zero at infinity

III. PROOF OF THEOREM 2

The proof will be subdivided into a few steps. In the first step the well known Rouché theorem will be extended to accommodate Hurwitz polynomials (a nonclosed Jordan curve, imaginary axis, will be considered). Then a simple approximation of d_+ will be provided. An elementary induction argument will conclude the proof.

Proposition 1 Extension of the Rouché' theorem for nonclosed Jordan curves (cf. [4], [8])

For a given Hurwitz polynomial $p_{n-1}(s) = p'_{n-1}(s)$ a perturbed polynomial

$$p_n(s) = p_{n-1}(s) + \delta(s) = (p'_{n-1} + \delta)S(s) \quad (9)$$

is Hurwitz as well if

$$\delta(jt)\delta(-jt) < p_{n-1}(jt)p_{n-1}(-jt) \quad t \geq 0 \quad (10)$$

Proof The proof follows directly from the zero exclusion principle. Consider a parametrized polynomial, cf. [4, Exercise 10 p. 5]

$$F(s, \lambda) = p_{n-1}(s) + \lambda\delta(s) \quad \lambda \in [0, 1] \quad (11)$$

In case of finite zeros of $F(s, \lambda)$ let us assume that condition (10) holds, but the polynomial

$$F(s, 1) = p_n(s) = p_{n-1}(s) + \delta(s) \quad (12)$$

is unstable. Due to continuity principle [8], [17] there exists $\lambda_0 \in [0, 1]$ and $t_0 \geq 0$ such that

$$F(jt_0, \lambda_0) = p_{n-1}(jt_0) + \lambda_0\delta(jt_0) = 0 \quad (13)$$

$$0 \leq |\lambda_0|^2 = \frac{|p_{n-1}(jt_0)|^2}{|\delta(jt_0)|^2} = \frac{p_{n-1}(jt_0)p_{n-1}(-jt_0)}{\delta(jt_0)\delta(-jt_0)} \leq 1 \quad (14)$$

and the latter condition contradicts (10). $F(s, \lambda)$ can also be destabilized through a zero at infinity. $F(s, \lambda)$ has a zero at infinity iff $sF(1/s, \lambda)$ has a zero at the origin. In this particular case there exists $\lambda_0 \in [0, 1]$ such that for $s_0 = 0$

$$s_0 F(1/s_0, \lambda_0) = s_0^n [p_{n-1}(1/s_0) + \lambda_0\delta(1/s_0)] = 0 \quad (15)$$

$$p_0^{n-1} + \lambda_0\delta_0 = 0 \quad \deg p_{n-1}(s) = \deg \delta(s) \quad (16)$$

where p_0^{n-1} and δ_0 are the highest term coefficients of $p_{n-1}(s)$ and $\delta(s)$

$$0 \leq |\lambda_0|^2 = \frac{|p_0^{n-1}|^2}{|\delta_0|^2} \leq 1 \quad (17)$$

Once again (17) contradicts (10) for $t_0 = 0$. If $\delta_0 = 0$, $|p_0^{n-1}|^2 > 0$ and (16) never happens. \square

To prove the main result of this note we shall need the following matrix

$$\Xi(jt) = S(jt)S'(-jt) \quad t \in (-\infty, \infty) \quad (18)$$

where $\sigma[\Xi(jt)]$ denotes the set of all eigenvalues of $\Xi(jt)$

Proposition 2

$$\Xi(jt) \text{ is a Hermitian matrix} \quad (19)$$

$$\text{rank } \Xi(jt) = 1 \quad t \in (-\infty, \infty) \quad (20)$$

$$\sigma[\Xi(jt)] = \{\lambda(t), 0, \dots, 0\} \quad (21)$$

$$\lambda(t) = \sum_{k=0}^n t^k = \begin{cases} \frac{1-t^{n+1}}{1-t} & t \neq \pm 1 \\ n+1 & t = \pm 1 \end{cases} \quad (22)$$

The corresponding eigenvectors are

$$\begin{aligned} v_1(jt) &= S(jt) \\ v_k(jt) &\in \ker S'(-jt) \\ k &= 1, \dots, n \end{aligned} \quad (23)$$

Proof It is easily verified that $\Xi(jt)$ is a Hermitian matrix of rank 1. Hence $\Xi(jt)$ has n eigenvalues equal to 0 and one nonzero eigenvalue $\lambda(t)$. Moreover, in this particular case

$$\lambda(t) = \text{tr } \Xi(jt)$$

and the proposition follows. \square

Proposition 3 If

$$\|\delta\| < [1/\lambda(t)]p_{n-1}(jt)p_{n-1}(-jt) \quad t \geq 0 \quad (24)$$

and the polynomial $p_{n-1}(s) = p'_{n-1}(s)$ is Hurwitz, then the polynomial

$$p_n(s) = (p'_{n-1} + \delta')S(s) \quad (25)$$

is Hurwitz as well.

Proof First let us observe, that according to Proposition 1

$$p_n(s) = (p'_{n-1} + \delta')S(s)$$

is stable if

$$p'_{n-1}\Xi(jt)p_{n-1} > \delta'\Xi(jt)\delta \quad t \geq 0 \quad (26)$$

According to the Rayleigh principle for Hermitian matrices [17, Section 6.2]

$$\delta'\Xi(jt)\delta \leq \lambda(t)\|\delta\|^2 \quad (27)$$

and we have the thesis of Proposition 3. \square

Proof of Theorem 2 From (24) we have (cf. [22])

$$r_{\min}^2(p_{nom}) \geq \inf_{t>0} \left[\frac{1}{\lambda(t)} p_{nom}(jt) p_{nom}(-jt) \right] \quad (28)$$

and in particular

$$r_{\omega}^2 \geq \inf_{t>0} \left[\frac{1}{\lambda(t)} p_{nom}(jt) p_{nom}(-jt) \right] \quad (29)$$

Our intention is to show that all Hurwitz polynomials with roots outside the unit circle (including circumference) and within the hyperbola (7) are destabilized by the zero at infinity, i.e., $r_{\min} = a_0^{1/m}$. The proof will use the induction argument with respect to the polynomial degree n . Coefficient "nom" notion will be omitted (case $n = 1$)

$$p_{nom}^1(s) = a_0 s + a_1 \quad (30)$$

If $s = -a_1/a_0$ satisfies $|s| \geq 1$ then $r_{\min} = \min(|a_0|, |a_1|) = |a_0|$

Case $n = 2$ $p_{nom}^2(s)$ is a nominal, Hurwitz polynomial of degree 2 and two cases can occur

1) $p_{nom}^2(s)$ has two complex, conjugate roots

2) $p_{nom}^2(s)$ has two real roots

In case 1)

$$p_{nom}^2(s) = a_0[(s + \alpha)^2 + \beta^2] \quad (31)$$

$$\begin{aligned} \frac{1}{\lambda_2(t)} p_{nom}^2(jt) p_{nom}^2(-jt) &= a_0^2 \frac{t^4 + 2t(\alpha^2 - \beta^2) + (\alpha^2 + \beta^2)t^2}{t^4 + t^2 + 1} \\ &\geq a_0^2 \frac{t^4 + t^2 + 1}{t^4 + t^2 + 1} = a_0^2 \end{aligned} \quad (32)$$

if $2(\alpha^2 - \beta^2) \geq 1$ ($\alpha^2 + \beta^2 \geq 1$)

Hence in view of (28)

$$r_{\min}^2 \geq a_0^2 \quad (33)$$

On the other hand, $r_{\min} \leq |a_0|$ and case 1) has been proved for complex roots. Case 2) is an immediate consequence of Case 1)

Assume that Theorem 2 holds for a polynomial $p_{nom}(s)$ of degree n we have to prove its validity for a polynomial of degree $n + 1$. First, let us assume that

$$p_{nom}^{n+1}(s) = p_{nom}(s)(s + \alpha), \quad \alpha \geq 1$$

$$\begin{aligned} \frac{1}{\lambda_{n+1}(t)} [p_{nom}^{n+1}(jt) p_{nom}^{n+1}(-jt)] &= \frac{p_{nom}(jt) p_{nom}(-jt) (t^2 + \alpha^2)}{1 + t^2 + \frac{t^{2(n+1)}}{t^{2(n+1)}}} \\ &= \frac{p_{nom}(jt) p_{nom}(-jt)}{1 + t^2 + \frac{t^{2n}}{t^{2(n+1)}}} \frac{1 + t^2 + \frac{t^{2n}}{t^{2(n+1)}}}{1 + t^2 + \frac{t^{2n}}{t^{2(n+1)}}} (t^2 + \alpha^2) \\ &\geq a_0^2 \frac{1 + t^2 + \frac{t^{2n}}{t^{2(n+1)}}}{1 + t^2 + \frac{t^{2n}}{t^{2(n+1)}}} (t^2 + \alpha^2) \geq a_0^2 \quad \alpha^2 \geq 1 \end{aligned} \quad (34)$$

If $p_{nom}^{n+2}(s) = p_{nom}^n(s)[(s + \alpha)^2 + \beta^2]$, then

$$\begin{aligned} \frac{1}{\lambda_{n+2}(t)} [p_{nom}^{n+2}(jt) p_{nom}^{n+2}(-jt)] &= \frac{p_{nom}^n(jt) p_{nom}^n(-jt)}{1 + t^2 + \frac{t^{2n}}{t^{2(n+2)}}} \frac{1 + t^2 + \frac{t^{2n}}{t^{2(n+2)}}}{1 + t^2 + \frac{t^{2n}}{t^{2(n+2)}}} \\ &\quad [t^4 + 2t^2(\alpha^2 - \beta^2) + (\alpha^2 + \beta^2)] \\ &\geq a_0^2 \end{aligned} \quad (35)$$

It is obvious, under assumptions of Theorem 2 that $a_0^2 \leq a_n^2$, and Theorem 2 is proved. \square

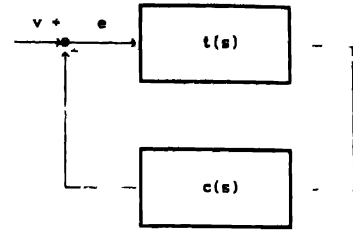


Fig. 2 A feedback control system

IV. ROBUST STABILIZING CONTROLLER DESIGN

A standard feedback stabilization problem is to design (for a proper, possibly unstable transfer function) a proper controller such that the denominator of the closed loop transfer function is stable

$$t(s) = p^{-1}(s)q(s) \quad (36)$$

$$p(s) = p_0 s^{n_1} + p_1 s^{n_1-1} + \dots + p_{n_1} \quad (37a)$$

$$q(s) = q_0 s^{n_2} + q_1 s^{n_2-1} + \dots + q_{n_2}, \quad n_1 \geq n_2 \quad (37b)$$

$$c(s) = a^{-1}(s)b(s) \quad (38)$$

$$a(s) = a_0 s^{m_1} + a_1 s^{m_1-1} + \dots + a_{m_1} \quad (39a)$$

$$b(s) = b_0 s^{m_2} + b_1 s^{m_2-1} + \dots + b_{m_2}, \quad m_1 \geq m_2 \quad (39b)$$

The transfer function of the closed loop system is

$$T_r(s) = \frac{a(s)q(s)}{p(s)a(s) + q(s)b(s)} = \frac{m(s)}{n(s)} \quad (40)$$

$$n(s) = n_0 s^{n_1+m_1} + \dots + n_{n_1+m_1} \quad (41a)$$

$$n_0 = \begin{cases} p_0 a_0 + q_0 b_0 & n_1 = n, m_1 = m \\ p_0 a_0 & \text{otherwise} \end{cases} \quad (41b)$$

Under very mild assumptions concerning $p(s)$ and $q(s)$ (p, q -coprime [3], [18]) a proper controller can always be determined such that roots of $n(s)$ are located arbitrarily in the left complex half plane [1], [3], [18]. If parameters of the plant are uncertain, for instance, cf. (37)

$$p^{min} \leq p \leq p^{max}, \quad q^{min} < q \leq q^{max} \quad (42)$$

it is quite reasonable to find a controller $c(s)$ (38) such that roots of the nominal polynomial $n_{nom}(s)$ (40)–(41) are located within appropriate region G of the left complex half plane and

$$n_{nom}(s) = p_{nom}(s)a(s) + q_{nom}(s)b(s) \quad (43)$$

a stability hypersphere of $n_{nom}(s)$ contains all uncertain polynomials $n(s)$ (40)–(42)

Usually, as a nominal polynomial $n_{nom}(s)$ a polynomial is selected such that its roots are all located within a specified region G of the left complex half plane, possibly contained within the hyperbola (7) and outside the unit circle centered at the origin. If $p_{nom}(s)$ and $q_{nom}(s)$ are coprime [3] a compensator $c(s) = a^{-1}(s)b(s)$ (38) can always be established, such that (43) holds

$$\begin{aligned} n_{nom}(s) &= n_0^{n_1+m_1} \prod_{i=1}^k (s + s_i) \prod_{j=1}^m [(s + \alpha_j)^2 + \beta_j^2] \\ n &= \deg n_{nom}(s) = 2m + k \end{aligned} \quad (44)$$

$$s_i \geq 1, s_i \in \mathcal{R}, i = 1, 2, \dots, k$$

$$\frac{\alpha_i^2}{\left(\frac{1}{\sqrt{2}}\right)^2} - \frac{\beta_i^2}{\left(\frac{1}{\sqrt{2}}\right)^2} \geq 1, \quad \alpha_i^2 + \beta_i^2 \geq 1,$$

$$i = 1, 2, \dots, m.$$

Theorem 3: Stability radius of $n_{\text{nom}}(s)$ (44) [cf. (40), (41) and (43)]

$$r_{\max}[n_{\text{nom}}(s)] = |n_0^{\text{nom}}|. \quad (45)$$

Proof: The proof is a direct consequence of Theorem 2. \square

Remark 2: Formulas (44) and (45) can be used to design proper, stabilizing robust feedback controllers for uncertain, single input single output systems (36). Parameters a_0, b_0 can be selected to provide an appropriate value of (45).

V. CONCLUSION AND DISCUSSION

This note presented a connection between pole placement in single input-single output systems and robust stability measure provided by real stability radius.

REFERENCES

- [1] S. P. Bhattacharyya, *Robust Stabilization Against Structured Perturbations* (Lecture Notes in Control and Information Sciences), vol. 99, New York: Springer-Verlag, 1987.
- [2] R. M. Biernacki, H. Hwang, and S. P. Bhattacharyya, "Robust stability with structured real parameter perturbations," *IEEE Trans. Automat. Contr.*, vol. AC-32, no. 6, 1987.
- [3] W. A. Wolovich, *Linear Multivariable Control*. New York: Springer-Verlag, 1974.
- [4] M. Marden, "The geometry of the zeros of a polynomial in a complex variable," *AMS Math. Surveys*, no. 3, 1949.
- [5] E. I. Jury, *Inners and Stability of Dynamic Systems*. New York: Wiley, 1974; see also Krieger, FL, 1982.
- [6] Y. C. Soh, et al., "Robust pole assignment," *Automatica*, vol. 23, no. 5, 1987.
- [7] J. Ackermann, "Parameter space design of robust control systems," *IEEE Trans. Automat. Contr.*, vol. AC-25, no. 6, 1980.
- [8] D. D. Siljak, *Nonlinear Systems, the Parameter Analysis and Design*. New York: Wiley, 1968.
- [9] M. K. Solak, "A note on robust G -stability," presented at the '90 Amer. Control Conf., San Diego, CA, May 23-25, 1990.
- [10] C. B. Soh, et al., "On the stability of polynomials with perturbed coefficients," *IEEE Trans. Automat. Contr.*, vol. AC-30, no. 10, 1985.
- [11] —, "Addendum to 'On the stability properties of polynomials with perturbed coefficients,'" *IEEE Trans. Automat. Contr.*, vol. AC-32, no. 3, 1987.
- [12] C. Soh, "On extending the hypersphere method to handle dominant pole assignment," *IEEE Trans. Automat. Contr.*, vol. AC-32, no. 3, 1987.
- [13] D. Hinrichsen and A. J. Pritchard, "Robustness measures for linear systems with application to stability radii of Hurwitz and Schur polynomials," *Int. J. Contr.*, vol. 55, no. 4, pp. 809-844, 1992.
- [14] D. Kaesbauer and J. Ackermann, "The distance from stability or G -stability boundaries," preprint of the 11th IFAC World Congress, Tallin, Estonia, USSR, vol. 5, pp. 130-134, Aug. 13-17, 1990.
- [15] K. P. Sondergeld, "A generalization of the Routh-Hurwitz stability criteria and an application to a problem in robust controller design," *IEEE Trans. Automat. Contr.*, vol. AC-28, pp. 965-970, Oct. 1983.
- [16] A. T. Fam and J. S. Meditch, "A canonical parameter space for linear system design," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 454-458, 1978.
- [17] J. N. Franklin, *Matrix Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1968.
- [18] C. T. Chen, *Linear System Theory and Design*. New York: Holt, Rinehart and Winston, 1988.

Pole Assignment for Uncertain Systems in a Specified Disk by State Feedback

Germain Garcia and Jacques Bernussou

Abstract—This paper presents a method for assigning the poles in a specified disk by state feedback for a linear discrete or continuous time uncertain system, the uncertainty being norm bounded. For this the "quadratic d stabilizability" concept which is the counterpart of quadratic stabilizability in the context of pole placement is defined and a necessary and sufficient condition for quadratic d stabilizability derived. This condition expressed as a parameter dependent discrete Riccati equation enables to design the control gain matrix by solving iteratively a discrete Riccati equation.

1. INTRODUCTION

Among the different ways for defining performance specifications, one of the most popular is to put some constraints on the poles location for the closed loop transfer matrix. In practice, the exact pole location is not required and it is often sufficient to locate them in a prescribed region in the left half complex plane for continuous time systems or in the unit disk for discrete time ones.

In this paper, the problem of pole assignment in a specified disk for linear discrete or continuous time uncertain system is addressed. In [1], the authors use a method based on conformal mapping which leads to a complex control law determination. In [2], another method working with the discrete Riccati equation for control law design is proposed. More recently, matricial algebraic tests for the spectrum of the state matrix to be clustered in a subregion of the complex plane were developed. In this direction, efforts to construct a modified Lyapunov matrix equation leading to efficient synthesis procedures have been made [3]–[6]. Unlike previous works, in the most recent works, uncertain systems are considered. A sufficient condition for disk location in the case of continuous or discrete systems is given in [5]. Robustness issues are discussed and margins for disk location are provided. In [4] and [3] conditions for pole placement in several regions (circle, vertical strip, sector...) are stated for uncertain systems, the parameter uncertainty being defined through convex bounded polytopes. The control law is derived via convex programming.

Here, following the same ideas as for quadratic stabilizability, a similar concept for pole assignment referred as "quadratic d stabilizability" is defined. First, a necessary and sufficient condition for quadratic d stabilizability is given for uncertainty affecting only the dynamical matrix. Expressed as a parameter dependent discrete Riccati equation, it leads to an algorithm allowing the state feedback gain design. The condition generalizes the sufficient condition given in [5]. The necessary and sufficient condition is extended to systems with uncertainty in both state and input matrices. This condition, although more involved, can be stated as a parameter dependent "augmented" Riccati equation. The paper ends with numerical examples illustrating the theory.

Manuscript received December 8, 1993; revised May 5, 1994.

G. Garcia is with the Laboratoire d'Automatique et d'Analyse des Systèmes du CNRS, 7, Avenue du Colonel Roche, 31077 Toulouse Cedex, France, and l'INSA, Complexe Scientifique de Rangueil, 31077 Toulouse Cedex, France.

J. Bernussou is with the Laboratoire d'Automatique et d'Analyse des Systèmes du CNRS, 7, Avenue du Colonel Roche, 31077 Toulouse Cedex, France.

IEEE Log Number 9406985.

II PRELIMINARIES

Let us consider a continuous or discrete time system described by

$$\delta[\dot{x}(t)] = (A + \Delta A)x(t) + Bu(t) \quad (1)$$

here $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $u(t) \in \mathbb{R}^m$ is the input and $x(t) \in \mathbb{R}^n$ is the state. δ is the derivation operator in the continuous time case [i.e., $\delta[x(t)] = \dot{x}(t)$] and the delay operator for the discrete time one [i.e., $\delta[x(t)] = x(t+1)$]. ΔA is the uncertainty of norm bounded type written as

$$\Delta A = DFF^T \quad (2)$$

where $D \in \mathbb{R}^{n \times n}$, $E \in \mathbb{R}^{n \times r}$ define the structure of the uncertainty and the parameter uncertainty F belongs to the set

$$\mathcal{K} = \{F \in \mathbb{R}^{r \times n} : F^T F \leq I\} \quad (3)$$

Throughout this paper the expression $(\bullet) \succ 0$ ($\prec 0$) means that matrix (\bullet) is positive definite (negative definite) and I denotes the identity matrix of appropriate dimension. The problem to be addressed is to determine the state feedback

$$u(t) = Kx(t) \quad (4)$$

such that the closed loop system poles lie in the disk $D(\alpha, r)$ with center $\alpha + j0$ and radius r . For the sake of simplicity let us introduce the following notations

$$\begin{aligned} A &= A + BK, & A &= (A - \alpha I)/r \\ A &= (A + BK - \alpha I)/r \\ \Delta A &= \Delta A/r \\ A_\Delta &= A + BK + \Delta A \end{aligned} \quad (5)$$

From the general result on matrix root clustering presented in [7] the following theorem can be deduced

Theorem 1 Let $A \in \mathbb{R}^{n \times n}$ be a given matrix. The eigenvalues of A belong to $D(\alpha, r)$ if and only if there exists a symmetric matrix $P \in \mathbb{R}^{n \times n}$ such that

$$\begin{pmatrix} -P^{-1} & A \\ A^T & -P \end{pmatrix} \prec 0 \quad (6)$$

Proof The proof is obtained from the standard Schur complement result and from [7].

From now on only the stable $D(\alpha, r)$ location cases are considered i.e. the one where $D(\alpha, r)$ belongs to the left half complex plane for continuous systems and to the unit disk for discrete systems. A concept related to Theorem 1 which follows the same philosophy is the quadratic stabilizability [8] can be introduced. This concept referred in the sequel as "quadratic d stabilizability" extends the condition of Theorem 1 to the case of uncertain system in a similar way as quadratic stabilizability extends the concept of stability to uncertain system.

Definition 1 The system (1) is a 'quadratically d stabilizable' by a linear state feedback $u(t) = Kx(t)$ if and only if there exists a symmetric matrix $P \in \mathbb{R}^{n \times n}$ such that

$$\forall \left(\begin{pmatrix} -P^{-1} & A + \Delta A \\ A^T + \Delta A^T & -P \end{pmatrix} \right) \prec 0 \quad (7)$$

for all $\Delta \in \mathbb{R}^{2n}$ and $F \in \mathcal{K}$

The above definition states that a system is quadratically d stabilizable if there exists a single matrix P satisfying (7) for all Δ in the uncertainty domain. It should be interesting to relate P to a Lyapunov matrix. The existence of a single Lyapunov matrix for all the systems in the uncertainty domain ensures the quadratic system stability for time varying uncertainties.

Theorem 2 If the system (1) is quadratically d stabilizable, it is quadratically stabilizable and P is a Lyapunov matrix for systems in the uncertainty domain.

Proof First, obviously $P \succ 0$. For continuous systems $\alpha = 0$ and $r < |\alpha|$. Developing (7) we obtain

$$(A + \Delta A)^T P + P(A + \Delta A) - \alpha^{-1}(A + \Delta A)P(\alpha I + \Delta A) + \alpha^{-1}(r - \alpha)I \prec 0$$

and then

$$(A + \Delta A)^T P + P(A + \Delta A) \prec \alpha^{-1}(A + \Delta A)^T P(\alpha I + \Delta A) - \alpha^{-1}(r - \alpha)P \prec 0$$

Thus the system is quadratically stabilizable.

For discrete systems the relation with quadratic stabilizability is not so straightforward. In fact the difficulty is due to the presence in the equation of the term

$$-\alpha[(\alpha I + \Delta A)^T P + P(\alpha I + \Delta A)]$$

Suppose that system (1) is quadratically d stabilizable. Using notations (7) is written

$$\begin{pmatrix} -P^{-1} & A_\Delta - \alpha I \\ A_\Delta - \alpha I & -P \end{pmatrix} \prec 0 \quad (8)$$

Let us first consider the case $\alpha > 0$. The circles in the unit disk are then characterized by $\alpha + r < 1$. Multiplying (8) on the left and on the right by the symmetric matrix

$$\begin{pmatrix} -\alpha P & \mathbf{1} \\ \mathbf{1} & P^{-1}/r \end{pmatrix}$$

we obtain

$$\begin{pmatrix} \bullet & \bullet \\ \bullet & T_1 \end{pmatrix} \prec 0$$

where T_1 satisfies $P T_1 P = -2P - 2\alpha P/r + 1/r(A_\Delta^T P + P A_\Delta) \prec 0$ with $r > 0$, $\alpha > 0$. This implies $-\alpha(A_\Delta^T P + P A_\Delta) > -2\alpha(\alpha + r)P$ and thus

$$\begin{aligned} A_\Delta^T P A_\Delta + (\alpha - r)P - 2\alpha(\alpha + r)P \\ = A_\Delta^T P A_\Delta - (\alpha + r)^2 P \\ < A_\Delta^T P A_\Delta + (\alpha^2 - r)P - \alpha(A_\Delta^T P + P A_\Delta) \\ < 0 \end{aligned}$$

But as $0 < \alpha + r < 1$ we conclude that

$$A_\Delta^T P A_\Delta - P < A_\Delta^T P A_\Delta - (\alpha + r)P < 0$$

The system is quadratically stabilizable and P is a Lyapunov matrix.

The case $\alpha < 0$ with $\alpha - r > -1$ is treated in much the same way as for $\alpha > 0$ beginning by multiplying (8) on the left and on the right by

$$\begin{pmatrix} -\alpha P & \mathbf{1} \\ \mathbf{1} & -P^{-1}/r \end{pmatrix} \quad (9)$$

III. MAIN RESULT

We are now in a position to give the main result of this paper. Denoting by

$$E_i = E/\sqrt{r}, \quad D_i = D/\sqrt{r}, \\ 4_{ii} = (A - a\mathbf{1})/r, \quad B_i = B/r$$

the following theorem can be stated.

Theorem 3: Let Q and R two positive definite symmetric matrices of appropriate dimension. The system (1) is quadratically d stabilizable by a linear state feedback $u(t) = Kx(t)$ if and only if there exist $\epsilon > 0$ and a positive definite symmetric matrix $P \in \mathcal{R}^{n \times n}$ satisfying the following discrete Riccati equation:

$$A'_{ii}(P^{-1} + B_i R^{-1} B'_i - \epsilon D_i D'_i)^{-1} A_{ii} \\ - P + \epsilon^{-1} E'_i E_i + Q = 0 \quad (9)$$

with

$$\epsilon^{-1} \mathbf{1} - D'_i P D_i > 0. \quad (10)$$

Then the control gain matrix K is given by

$$K = -R^{-1} B'_i (P^{-1} + B_i R^{-1} B'_i - \epsilon D_i D'_i)^{-1} A_{ii}. \quad (11)$$

Remark. Let $P = \epsilon P$ and using the matrix inversion lemma, the Riccati equation (9) can be written in a more "classical form":

$$A'_{ii} \hat{P} A_{ii} - \hat{P} - A'_{ii} \hat{P} (B_i - D_i) \\ \cdot \begin{pmatrix} \epsilon R + B'_i \hat{P} B_i & B'_i \hat{P} D_i \\ D'_i \hat{P} B_i & D'_i \hat{P} D_i - \mathbf{1} \end{pmatrix}^{-1} \begin{pmatrix} B'_i \\ D'_i \end{pmatrix} \hat{P} A_{ii} \\ + E'_i E_i + \epsilon Q = 0.$$

The Riccati equation (9) is of a type which also arises in linear quadratic differential games (discrete time case) in which the cost functional is (see [9])

$$J_i(u_k, v_k) = \sum_{k=0}^{\infty} [y'_k y_k + \epsilon x'_k Q x_k + \epsilon u'_k R u_k - v'_k v_k]$$

where u_k is the minimizing control, v_k is the maximizing control, and the system under consideration is described by state equation

$$x_{k+1} = A_{ii} x_k + B_i u_k + D_i v_k \\ y_k = E_i x_k.$$

Proof of Theorem 3 For the proof, some technical lemmas are necessary. In what follows, we give almost all them without proofs which can be found in the quoted references.

Lemma 1 [10] Let S, Y , and Z be given by $k \times k$ symmetric matrices such that $S \geq 0$, $Y < 0$, and $Z \geq 0$. Furthermore, assume that

$$(\eta' Y \eta)^2 - 4(\eta' S \eta)(\eta' Z \eta) > 0$$

for all nonzero $\eta \in \mathcal{R}^k$. Then, there exists a constant $\lambda > 0$ such that the matrix $\lambda^2 S + \lambda Y + Z$ is negative definite. •

Lemma 2 (Finsler Lemma). Let X be a $k \times k$ symmetric matrix and $B_i \in \mathcal{R}^{k \times m}$ such that $\eta' X \eta < 0$ for all $\eta \neq 0$ satisfying $B'_i \eta = 0$. Then there exists a positive definite symmetric matrix $Y \in \mathcal{R}^{k \times k}$ such that the matrix $X - B_i Y B'_i$ is negative definite. •

Lemma 3 [10], [11] Given $x \in \mathcal{R}^n$, $y \in \mathcal{R}^m$,

$$\text{Max} \{ (x' D_i F E_i y) : F' F \leq \mathbf{1} \} = x' D_i D'_i x y' E'_i E_i y. \quad (12)$$

Lemma 4 [11] Let Q and R be two positive symmetric matrices such that the Riccati equation (9) has positive definite symmetric solution for $\epsilon^* > 0$. Then for any given positive definite symmetric matrices \hat{Q} and \hat{R} , there exists $\hat{\epsilon} > 0$ such that (9) has a definite positive symmetric solution for all $\epsilon \in (0, \hat{\epsilon})$ with R replaced by \hat{R} and Q by \hat{Q} . •

The above lemma states that the existence of a matrix P satisfying the conditions of Theorem 3 does not depend upon the choice of the R and Q matrices.

Lemma 5: Let D_i, A_i, E_i be matrices of appropriate dimension. Let $F \in \mathcal{N}$ and P be a positive definite symmetric matrix satisfying

$$\epsilon^{-1} \mathbf{1} - D'_i P D_i > 0, \quad \epsilon > 0.$$

Then,

$$A' P D_i F E_i + E'_i F' D'_i P A_i + E'_i F' D'_i P D_i F E_i \\ \leq A' P D_i (\epsilon^{-1} \mathbf{1} - D'_i P D_i)^{-1} D'_i P A_i + \epsilon^{-1} F'_i E_i. \quad \bullet$$

Proof If we define

$$Y = (\epsilon^{-1} \mathbf{1} - D'_i P D_i)^{-1/2} D'_i P A_i - (\epsilon^{-1} \mathbf{1} - D'_i P D_i)^{1/2} F E_i.$$

We have $Y' Y \geq 0$ and then

$$A' P D_i (\epsilon^{-1} \mathbf{1} - D'_i P D_i)^{-1} D'_i P A_i - A' P D_i F E_i \\ - E'_i F' D'_i P A_i + E'_i F' (\epsilon^{-1} \mathbf{1} - D'_i P D_i) F F_i \geq 0.$$

Rearranging some terms, the lemma is proved. □

Necessity: Suppose the system (1) quadratically d stabilizable by a state feedback $u(t) = Kx(t)$. Then, there exists a positive definite symmetric matrix $P \in \mathcal{R}^{n \times n}$ such that for all $X \in \mathcal{R}^{2n}$ and $F \in \mathcal{N}$:

$$X' \begin{pmatrix} -P^{-1} & A_{ii} + D_i F E_i + B_i K \\ A'_{ii} & -P \end{pmatrix} X < 0$$

which can be written

$$X' \begin{pmatrix} -P^{-1} & A_{ii} \\ A'_{ii} & -P \end{pmatrix} X + X' \begin{pmatrix} 0 & B_i K \\ K' B'_i & 0 \end{pmatrix} X \\ + X' \begin{pmatrix} 0 & D_i F E_i \\ E'_i F' D'_i & 0 \end{pmatrix} X < 0. \quad (13)$$

And then for all $X \neq 0$ such that $[B'_i \ 0]X = 0$, we have

$$X' \begin{pmatrix} -P^{-1} & A_{ii} \\ A'_{ii} & -P \end{pmatrix} X + X' \begin{pmatrix} 0 & D_i F E_i \\ E'_i F' D'_i & 0 \end{pmatrix} X < 0.$$

This implies that

$$X' \begin{pmatrix} -P^{-1} & A_{ii} \\ A'_{ii} & -P \end{pmatrix} X \\ < - \left\{ \text{Max } X' \begin{pmatrix} 0 & D_i F E_i \\ E'_i F' D'_i & 0 \end{pmatrix} X : F' F \leq \mathbf{1} \right\} \leq 0$$

for all $X \neq 0$ such that $[B'_i \ 0]X = 0$. And hence

$$\left(X' \begin{pmatrix} -P^{-1} & A_{ii} \\ A'_{ii} & -P \end{pmatrix} X \right)^2 \\ > \left\{ \text{Max } X' \begin{pmatrix} 0 & D_i F E_i \\ E'_i F' D'_i & 0 \end{pmatrix} X : F' F \leq \mathbf{1} \right\}^2$$

for all $X \neq 0$ such that $[B'_i \ 0]X = 0$. Writing $X = [x' y']'$ and by Lemma 3 we obtain

$$\left(X' \begin{pmatrix} -P^{-1} & A_{ii} \\ A'_{ii} & -P \end{pmatrix} X \right)^2 \\ > 4 X' \begin{pmatrix} D_i D'_i & 0 \\ 0 & 0 \end{pmatrix} X X' \begin{pmatrix} 0 & 0 \\ 0 & E'_i E_i \end{pmatrix} X \quad (14)$$

all $\lambda \neq 0$ such that $[B' \ 0]\lambda = 0$. Now denoting by Γ a matrix whose columns form a set of basis vector for the linear space $\in \mathcal{R}^{2n}$ $[B' \ 0]\lambda = 0$ and defining

$$\begin{aligned} \zeta &= \Gamma' \begin{pmatrix} D & D' & 0 \\ 0 & 0 & 0 \end{pmatrix} \Gamma \geq 0 \\ \gamma &= \Gamma' \begin{pmatrix} -P^{-1} & 4 \\ 4' & -P \end{pmatrix} \Gamma < 0 \\ \mathcal{Z} &= \Gamma' \begin{pmatrix} 0 & 0 \\ 0 & F'F \end{pmatrix} \Gamma \geq 0 \end{aligned}$$

It follows by Lemma 1 that there exists a constant $\epsilon > 0$ such that

$$\begin{aligned} \lambda \begin{pmatrix} D & D' & 0 \\ 0 & 0 & 0 \end{pmatrix} \lambda + \lambda \begin{pmatrix} -P^{-1} & 4 \\ 4' & -P \end{pmatrix} \lambda \\ + \frac{1}{\epsilon} \lambda \begin{pmatrix} 0 & 0 \\ 0 & F'F \end{pmatrix} \lambda < 0 \end{aligned}$$

for all $\lambda \neq 0$ such that $[B' \ 0]\lambda = 0$. By Lemma 2 there exists a positive definite symmetric matrix $W = R^{-1}$ such that the matrix

$$\begin{pmatrix} -P^{-1} + \epsilon D D' & 4 \\ 4' & -P + \epsilon F'F \end{pmatrix} = \begin{pmatrix} B R^{-1} B & 0 \\ 0 & 0 \end{pmatrix}$$

is negative definite. P being positive definite by hypothesis. This implies

$$4 (P^{-1} + B R^{-1} B - \epsilon D D')^{-1} 4 - P + \epsilon^{-1} F'F < 0$$

And there exists $Q = Q' > 0$ such that

$$4 (P^{-1} + B R^{-1} B - \epsilon D D')^{-1} 4 - P + \epsilon^{-1} F'F < -Q < 0$$

It remains to show the following inequality

$$\epsilon^{-1} \mathbf{1} - D P D' > 0 \quad (15)$$

Following the same development as above we obtain

$$\begin{aligned} \epsilon \lambda \begin{pmatrix} D & D' & 0 \\ 0 & 0 & 0 \end{pmatrix} \lambda + \lambda \begin{pmatrix} P^{-1} & 4 \\ 4' & -P \end{pmatrix} \lambda \\ + \frac{1}{\epsilon} \lambda \begin{pmatrix} 0 & 0 \\ 0 & F'F \end{pmatrix} \lambda < 0 \end{aligned}$$

for all $\lambda \neq 0$ such that $[0 \ B \ K]\lambda = 0$. Now from Lemma 2 there exists a matrix $V = V' > 0$ such that

$$\begin{pmatrix} -P^{-1} + \epsilon D D' & 4 \\ 4' & -P + \epsilon F'F \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & K B V B K \end{pmatrix} < 0$$

which implies that $P^{-1} - \epsilon D D' > 0$ or equivalently $\epsilon^{-1} \mathbf{1} - D P D' > 0$

Sufficiency Suppose that condition of Theorem 3 is satisfied and let

$$\mathcal{L} = (4' + \Delta 4') P (4' + \Delta 4') - P$$

Using Lemma 5 and (5), we have

$$\begin{aligned} \mathcal{L} &\leq [4' \ P 4' - P + 4' P D (\epsilon^{-1} \mathbf{1} - D' P D)^{-1} D' P 4' \\ &\quad + \epsilon^{-1} L' F] \\ &= [4' (P^{-1} - \epsilon D D')^{-1} 4' - P + \epsilon^{-1} F'F] \end{aligned} \quad (16)$$

By the conditions of Theorem 3

$$P = 4' (P^{-1} + B R^{-1} B' - \epsilon D D')^{-1} 4' + \epsilon^{-1} I' L + Q$$

Replacing in (16) leads to

$$\begin{aligned} \mathcal{L} &\leq [4' (P^{-1} - \epsilon D D')^{-1} 4' \\ &\quad - 4' (P^{-1} + B R^{-1} B' - \epsilon D D')^{-1} 4' - Q] \end{aligned} \quad (17)$$

Now denoting $\Gamma = (P^{-1} + B R^{-1} B' - \epsilon D D')^{-1} 4'$ that

$$\begin{aligned} 4' (P^{-1} + B R^{-1} B' - \epsilon D D')^{-1} 4' \\ = \Gamma' B R^{-1} B' \Gamma + I^{-1} \end{aligned}$$

(17) is now given by

$$\begin{aligned} \mathcal{L} &= [4' (P^{-1} - \epsilon D D')^{-1} 4' - Q] \\ &\quad - \Gamma' B R^{-1} B' \Gamma - I^{-1} (P^{-1} - \epsilon D D')^{-1} 4' \end{aligned}$$

A simple calculation shows that with K given by (11)

$$4' (P^{-1} - \epsilon D D')^{-1} 4' = \Gamma' (P^{-1} - \epsilon D D')^{-1} I = 0$$

and then

$$\mathcal{L} < -[Q + K' R K] < -Q < 0$$

Then the system is quadratically d stabilizable and the proof is completed \square

A d Stabilization Algorithm

Theorem 3 gives a condition which allows to derive the d stabilization algorithm presented below

- Step 1 Choose any positive definite symmetric weighting matrices $Q \in \mathcal{R}^{n \times n}$ and $R \in \mathcal{R}^{m \times m}$ and initialize ϵ to some starting value
- Step 2 Determine whether the Riccati equation (9) has positive definite solution satisfying (10). If such a solution exists the algorithm succeeds and a control law is determined by (11). If not go to step 3
- Step 3 Take $\epsilon = \epsilon/2$. If ϵ is less than some computational accuracy ϵ stop the algorithm fails. Otherwise go to step 2

Remark In order to solve the Riccati equation in step 2 some standard methods can be used see for example [12] [13]

B Comparison to a Previous Result

In this paragraph we examine the relation between the results given in this paper and the results in [5]. First suppose that system (1) is well known (i.e. $D = 0$ and $I = 0$) that is

$$\delta[x(t)] = Ax(t) + Bu(t) \quad (18)$$

Following the same steps as for the proof of Theorem 3 the result below can be proved

Corollary 1 Let Q and R be two positive definite symmetric matrices of appropriate dimension. The system (18) is d stabilizable by a linear state feedback $u(t) = Kx(t)$ (i.e. all poles of the closed loop system lie in $D(\alpha)$) if and only if there exists a positive definite symmetric matrix $P' \in \mathcal{R}^{n \times n}$ satisfying the following discrete Riccati equation

$$4 (P^{-1} + B R^{-1} B')^{-1} 4' = P + Q < 0 \quad (19)$$

Then the control gain matrix K is given by

$$\begin{aligned} K &= -R^{-1} B' (P^{-1} + B R^{-1} B')^{-1} 4' \\ &= -(I^* R + B' P B)^{-1} B' P (4 - \alpha \mathbf{1}) \end{aligned} \quad (20)$$

Thus the sufficient condition stated in [5] is also necessary

IV UNCERTAINTY IN BOTH THE STATE AND INPUT MATRICES

In this section, the problem of uncertainty on both state and input matrices is investigated. Let the system

$$\dot{x}(t) = (A + \Delta A)x(t) + (B + \Delta B)u(t) \quad (21)$$

with

$$[\Delta A \quad \Delta B] = DF[\Gamma_1 \quad \Gamma_2]$$

We begin by defining an "extended system" [14]

$$\dot{X}(t) = (F + \Delta F)X(t) + B_F u(t) \quad (22)$$

$$F = \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix}$$

$$B_F = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \Delta F = D_F F E_F \quad D_F = \begin{pmatrix} D/\sqrt{\gamma} \\ 0 \end{pmatrix}$$

$$\Delta B = \Delta B/\gamma \quad E_F = (E_1/\sqrt{\gamma} \quad E_2/\sqrt{\gamma}) \quad \rho = n + m$$

Theorem 4 Let $Q_F \in \mathbb{R}^{(n+m) \times (n+m)}$ and $R_F \in \mathbb{R}^{m \times m}$ be positive definite symmetric matrices. The system (21) is quadratically d stabilizable by a linear state feedback $u(t) = Kx(t)$ if and only if there exist $\eta > 0$ and a positive definite symmetric matrix $\Pi \in \mathbb{R}^{(n+m) \times (n+m)}$ satisfying the following discrete Riccati equation

$$F'(\Pi^{-1} + B_F R_F^{-1} B_F' - \eta D_F D_F')^{-1} F - \Pi + \eta^{-1} F' F + Q_F = 0 \quad (23)$$

with

$$\eta^{-1} \mathbf{1} - D_F' S D_F > 0 \quad (24)$$

The control gain matrix is then given by

$$u = S^{-1} x_1 \quad (25)$$

where $S = \Pi^{-1}$ is partitioned as follows

$$S = \begin{pmatrix} S_1 & S_2 \\ S_2' & S_3 \end{pmatrix} \quad S_1 \in \mathbb{R}^{n \times n} \quad S_3 \in \mathbb{R}^{m \times m} \quad \bullet$$

For the proof we need the following lemmas

Lemma 6 Let a d stable system (i.e. all its modes are in $D(n, \gamma)$) be given by

$$\dot{x}(t) = Ax(t) \quad (26)$$

Thus there exists a definite positive symmetric matrix Γ such that

$$\Gamma \Gamma' - \Gamma < 0$$

Then

$$A \Gamma \Gamma' - \Gamma < 0$$

is satisfied with $\Pi = \Gamma^{-1}$ •

Lemma 7 The system (21) is quadratically d stabilizable if and only if there exists $\mathcal{W} = \mathcal{W}' > 0$ such that

$$(1 \quad 0)[(F + \Delta F)\mathcal{W}(F + \Delta F)' - \mathcal{W}]\begin{pmatrix} 1 \\ 0 \end{pmatrix} < 0 \quad \bullet$$

Proof The system (21) is quadratically d stabilizable by a linear state feedback $u(t) = Kx(t)$ if and only if there exists a positive definite symmetric matrix P such that

$$[(1 \quad 0 + \Delta A/\gamma) + (B/\gamma + \Delta B/\gamma)K]P[(1 \quad 0 + \Delta A/\gamma)' + (B/\gamma + \Delta B/\gamma)K]' - P < 0$$

This is equivalent to

$$(1' \quad 0) \left[(F + \Delta F) \begin{pmatrix} P & PK' \\ K'P & K'PK' \end{pmatrix} (F + \Delta F)' - \begin{pmatrix} P & PK' \\ K'P & K'PK' \end{pmatrix} \right] \begin{pmatrix} 1 \\ 0 \end{pmatrix} < 0$$

and then there exists $\rho > 0$ (sufficiently small) such that

$$(1' \quad 0) \left[(F + \Delta F) \underbrace{\begin{pmatrix} P & PK' \\ K'P & K'PK' + \rho \mathbf{1} \end{pmatrix}}_W (F + \Delta F)' - \underbrace{\begin{pmatrix} P & PK' \\ K'P & K'PK' + \rho \mathbf{1} \end{pmatrix}}_W \right] \begin{pmatrix} 1 \\ 0 \end{pmatrix} < 0$$

For the converse see [15]

Proof of Theorem 4 From Lemma 7 the necessity is derived in a similar way as for Theorem 3. We prove only the sufficiency. Suppose that the condition of theorem is satisfied and let

$$H = [A + \Delta A \quad (B + \Delta B)S^{-1}] S_1 [A + \Delta A \quad (B + \Delta B)S^{-1}]' - S_1$$

Developing this expression leads to

$$H = (A + \Delta A)S_1(A + \Delta A)' - S_1 + (B + \Delta B)S_2(1 + \Delta A)' + (A + \Delta A)S_2(B + \Delta B) + (B + \Delta B)S_2' S_1^{-1} S_2(B + \Delta B)$$

But by definite positivity of S we have

$$S_1 - S S_1^{-1} S_2 > 0$$

and then

$$H \leq (1 \quad 0) \left[\begin{pmatrix} A + \Delta A & B + \Delta B \\ 0 & 0 \end{pmatrix} \begin{pmatrix} S_1 & S_2 \\ S_2' & S_3 \end{pmatrix} \begin{pmatrix} A + \Delta A & B + \Delta B \\ 0 & 0 \end{pmatrix}' - \begin{pmatrix} S_1 & S_2 \\ S_2' & S_3 \end{pmatrix} \right] \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

or

$$H < (1 \quad 0)[(F + \Delta F)S(F + \Delta F)' - S](1 \quad 0)'$$

By Lemma 5 we have

$$H \leq (1 \quad 0)[F(S^{-1} - \eta^{-1}F'F'F)^{-1}F' - S + \eta D_F D_F'](1 \quad 0)'$$

with by (23)

$$\eta \mathbf{1} - F' S F' > 0$$

Because of the structure of B_F and by (23) it is now easy to see that the condition above can be written as follows

$$H \leq (1 \quad 0)[F(S^{-1} - \eta^{-1}F'F'F)^{-1}F' - S - B_F R_F^{-1} B_F' + \eta D_F D_F'](1 \quad 0)' < 0$$

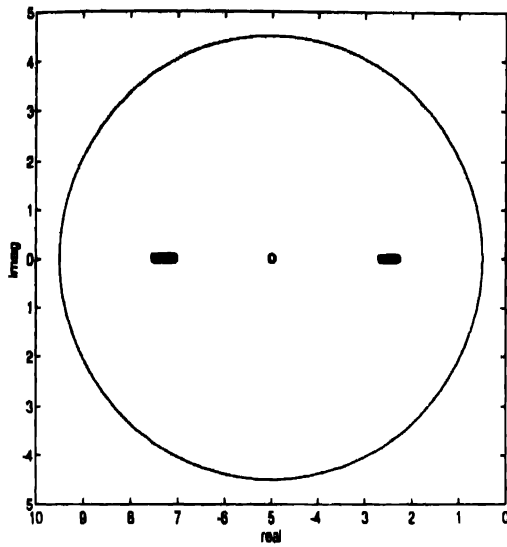


Fig. 1 Closed loop modes

and it follows that the system

$$\delta[\dot{x}(t)] = [(A + \Delta A) + (B + \Delta B)S_1^{-1}]x(t)$$

is quadratically d stabilizable for all uncertainty in the uncertainty domain. By Lemma 6, the system

$$\delta[\dot{x}(t)] = [(A + \Delta A) + (B + \Delta B)S_1^{-1}]x(t)$$

is quadratically d stabilizable over the uncertainty domain. \square

V. NUMERICAL EXAMPLES

A Continuous System [16] Let the continuous time uncertain system

$$\dot{x}(t) = (A + \Delta A)x(t) + Bu(t) \quad (27)$$

$$A = \begin{pmatrix} 0.52 & 17.76 & 90.21 \\ 0.17 & -0.75 & 11.10 \\ 0 & 0 & -2.50 \end{pmatrix} \quad B = \begin{pmatrix} -0.124 \\ 0 \\ 2.50 \end{pmatrix}$$

The uncertainty structure is described by

$$D = \begin{pmatrix} 0.5 \\ -0.25 \\ 0 \end{pmatrix} \quad F = [0.10 \quad 0.10 \quad 1.00] \quad I = f \in \mathfrak{R}$$

We want to place the poles in a disk centered at $\alpha = -5$ and with a radius $r = 4.5$. This problem was treated in [4] for convex bounded uncertainties. The Q and R matrices are chosen to be

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad R = 1$$

The algorithm is initialized with $\epsilon = 10$. After 3 iterations, for $\epsilon = 1.25$

$$P = \begin{pmatrix} 3.219 & 13.462 & -11.887 \\ 13.462 & 89.288 & -95.452 \\ -11.887 & -95.452 & 111.425 \end{pmatrix}$$

$$\epsilon^{-1} \mathbf{1} - D'PD = 0.295$$

The control gain matrix is then given by

$$K = [0.0016 \quad 0.0173 \quad 0.9178]$$

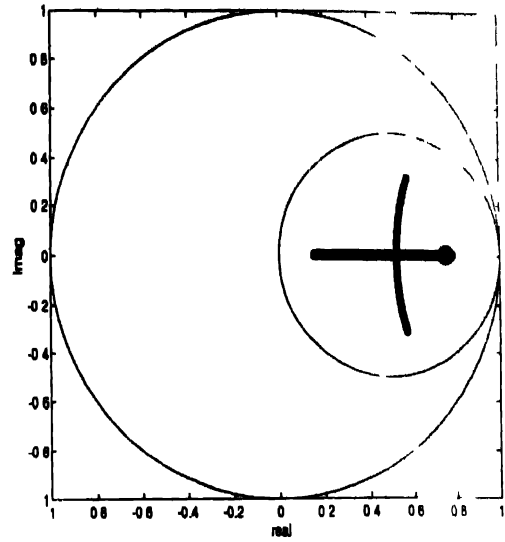


Fig. 2 Closed loop modes

Fig. 1 shows the closed loop modes for 200 values of t such that $-1 \leq f \leq 1$.

A Discrete System Let the discrete time uncertain system

$$x(t+1) = (A + \Delta A)x(t) + Bu(t) \quad (28)$$

$$A = \begin{pmatrix} -0.50 & 1.00 & 1.00 \\ 1.25 & -0.25 & -1.00 \\ -2.25 & -0.25 & 0.50 \end{pmatrix} \quad B = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

The uncertainty structure is described by

$$D = \begin{pmatrix} 0.3333 \\ 0.1667 \\ 0.1667 \end{pmatrix} \quad F = [0.1944 \quad 0.0833 \quad 0.1389]$$

$$I = f \in \mathfrak{R}$$

This nominal system has an uncontrollable stable mode at 0.75. We try to place the poles in the circle centered at $\alpha = 0$ with a radius $r = 0.5$. For the Q and R matrices, we take

$$Q = 1 \quad R = 1$$

The algorithm is initialized with $\epsilon = 10$. After 7 iterations, $\epsilon = 0.0781$ and

$$P = \begin{pmatrix} 50.151 & 2.138 & -0.351 \\ 2.138 & 2.916 & 1.409 \\ -0.351 & 1.309 & 2.491 \end{pmatrix}$$

$$\epsilon^{-1} \mathbf{1} - D'PD = 0.514$$

The control gain matrix is then given by

$$K = [1.0421 \quad -1.0133 \quad -0.9996]$$

Fig. 2 shows the closed loop modes for 200 values of t such that $-1 \leq f \leq 1$.

VI. CONCLUSION

In this paper, pole assignment in a specified disk for discrete or continuous time uncertain systems has been investigated. The condition derived is expressed as a parameter dependent discrete Riccati equation leading to a simple procedure for the design of the state feedback gain matrix. Numerical examples for continuous and discrete time systems have been presented showing the efficiency of the proposed method. The results of this paper are based on the strong assumption of state availability. In practice, for most systems, only partial state being available the problem of pole assignment by output feedback is of great interest. This problem will be addressed in the near future.

ACKNOWLEDGMENT

We would like to acknowledge Reviewer 1 for helpful comments on the numerical examples.

REFERENCES

- [1] A. V. Bogachev, V. V. Grigorev, V. N. Drozdov, and A. N. Korovyakov, "Analytic design of controls from root indicators," *Automat. Remote Contr.*, vol. 40, no. 8, 1979.
- [2] Y. Mori and Y. Shimemura, "On a design method for feedback control law to locate eigenvalues in a specified region," *S.I.C.E.*, vol. 16, no. 3, 1980 (in Japanese).
- [3] D. Arzelier, J. Bernussou, and G. Garcia, "Pole assignment of linear uncertain systems in a sector via a Lyapunov-type approach," *IEEE Trans. Automat. Contr.*, vol. 38, no. 7, 1993.
- [4] J. C. Geromel, G. Garcia, and J. Bernussou, "H² robust control with pole placement," in *Proc. 12th World I.F.A.C. Congress*, Sydney, Australia, 1993.
- [5] K. Furuta, and S. B. Kim, "Pole assignment in a specified disk," *IEEE Trans. Automat. Contr.*, vol. 32, no. 5, 1987.
- [6] W. M. Haddad, D. S. Bernstein, "Controller design with regional pole constraints," *IEEE Trans. Automat. Contr.*, vol. 37, no. 1, 1992.
- [7] A. G. Mazco, "The Lyapunov matrix equation for a certain class of regions bounded by algebraic curves," *Soviet. Automat. Contr.*, 1980.
- [8] B. R. Barmish, "Necessary and sufficient conditions for quadratic stabilizability of an uncertain system," *J. Opt. Theory Appl.*, vol. 46, no. 4, 1985.
- [9] T. Basar and P. Bernhard, *H[∞]-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Boston, MA: Birkhäuser, 1991.
- [10] I. R. Petersen, "A stabilization algorithm for a class of uncertain linear systems," *Syst. Contr. Lett.*, vol. 8, 1987.
- [11] G. Garcia, J. Bernussou, and D. Arzelier, "Robust stabilization of discrete time linear systems with norm bounded time varying uncertainty," *Syst. Contr. Lett.*, 1993.
- [12] S. Bittanti, A. J. Laub, and J. Willems, Eds., *The Riccati Equation*. New York: Springer-Verlag, 1991.
- [13] T. Pappas, A. J. Laub, and N. R. Sandell, "On the numerical solution to the discrete-time algebraic Riccati equation," *IEEE Trans. Automat. Contr.*, vol. 25, no. 4, 1980.
- [14] B. R. Barmish, "Stabilization of uncertain systems via linear control," *IEEE Trans. Automat. Contr.*, vol. 28, no. 8, 1983.
- [15] J. C. Geromel, P. L. D. Peres, and J. Bernussou, "On a convex parameter space method for linear control design of uncertain systems," *SIAM Contr. Opt.*, vol. 29, no. 2, 1991.
- [16] W. E. Schmitendorf, "Designing stabilizing controllers for uncertain systems using the Riccati equation approach," *IEEE Trans. Automat. Contr.*, vol. 33, no. 4, 1988.

Modified Output Error Identification—Elimination of the SPR Condition

Amir Betser and Ezra Zeheb

Abstract—In this note, we present a modified adaptive output error algorithm for identification. We refer to both the continuous time and the discrete time cases. In the standard algorithm, stability requires that a certain transfer function should be Strictly Positive Real (SPR); however, in the present algorithm, the SPR condition, which is difficult to satisfy, is eliminated. This is achieved by adding a fixed feedback gain. When there exists an *a priori* partial knowledge about the parameters of the plant, we provide a simple procedure to design that gain. An illustrative example comparing the modified algorithm to the standard one is provided.

I. INTRODUCTION

A sufficient condition for asymptotic stability of the adaptive output error algorithm for identification is that a certain transfer function should be a Strictly Positive Real (SPR) rational function [1]–[3]. A drawback of the algorithm is the need to determine the SPR function. This is a very difficult task which has not yet been satisfactorily solved. Although the numerator of that function is an arbitrary polynomial which is chosen by the designer of the system, the denominator of the function must be the denominator of the transfer function of the plant, which is either unknown or only partially known.

Several papers give necessary and sufficient conditions to check whether a transfer function, with uncertain parameters, is a SPR transfer function [4]–[8]. Some papers treated the design problem, where there exists an *a priori* knowledge about the plant, i.e., it is assumed that the denominator $a(s)$ of the transfer function of the plant belongs to a certain set of stable polynomials. In the continuous time case, only sufficient conditions were found for the existence of a fixed polynomial $g(s)$ such that $g(s)/a(s)$ is an SPR transfer function, where $a(s)$ is the above-mentioned denominator [4], [9]–[11]. As mentioned above, the design of such a polynomial $g(s)$ is (for degrees higher than 3), difficult and not yet satisfactorily solved even when there is an *a priori* knowledge about the plant. In the discrete time case, necessary and sufficient conditions were found for the existence of a fixed polynomial $c(z^{-1})$ such that $c(z^{-1})/a(z^{-1})$ is a Discrete SPR transfer function where $a(z^{-1})$ is the denominator of the transfer function of the plant, which belongs to a certain stable set of polynomials [9].

In this note, we present a modified output error algorithm, pertaining to both the continuous time and the discrete time cases. The modification includes a fixed gain feedback which, if designed properly, eliminates the SPR requirement.

The modified algorithms are presented in Sections II and III for the continuous time and the discrete time cases, respectively. In Section IV, we use the modified algorithms to solve the design problem, where the denominator of the plant belongs to the same set of stable polynomials mentioned above. In Section V, we give an illustrative example where the advantages of the modified algorithm are evident. Section VI is the conclusion.

Manuscript received January, 1993; revised March, 1994. This work was supported in part by the Fund for the Promotion of Research at the Technion. The authors are with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Technion City, Haifa 32000, Israel. IEEE Log Number 9406990.

II MODIFIED ALGORITHM—CONTINUOUS TIME CASE

We start by following the derivation described in [2] of the output error algorithm. Consider the identification problem of a single input single output, linear time invariant system described by the transfer function

$$P(s) = \frac{b(s)}{a(s)} = \frac{\sum_{i=0}^m b_i s^i}{s^n + \sum_{i=1}^n a_i s^i} \quad m < n \quad (1)$$

Denote the input and the output of the system by $u(t)$ and $y(t)$, respectively. We assume that $a(s)$ is a stable polynomial (i.e., all the zeros of $a(s)$ are in the open left half plane). Let $F_i(s) \triangleq s^i / q(s)$, $i = 1, 2, \dots, n$ where $q(s) = s^n + \sum_{i=1}^n q_i s^i$ is a stable arbitrary polynomial. The input-output relationship can be expressed as follows

$$y(t) = c_0^T(t) \theta_0 \quad (2)$$

where

$$c_0^T(t) = [F_1 u \quad F_2 u \quad \dots \quad F_n u \quad -F_1 y \quad \dots \quad -F_n y] \quad (3)$$

$$\theta_0^T = [b_0 \quad b_1 \quad \dots \quad b_m \quad a_1 - q_1 \quad \dots \quad a_n - q_n] \quad (4)$$

and $F_i u$, $F_i y$ are the outputs of the stable filters $F_i(s)$ with the inputs $u(t)$, $y(t)$ respectively. At this point, we depart from the standard output error algorithm. We define our output error as usually done by

$$e(t) = y(t) - \hat{y}(t) \quad (5)$$

However, our $\hat{y}(t)$ differs from the standard definition. We add a feedback gain as follows

$$\hat{y}(t) = c^T(t) \theta(t) + l_1(t) \quad l_1(t) = 0 \quad (6)$$

where

$$c^T(t) = [F_1 u \quad F_2 u \quad \dots \quad F_n u \quad -F_1 y \quad \dots \quad -F_n y] \quad (7)$$

$$\theta^T(t) = [b_0 \quad b_1 \quad \dots \quad b_m \quad a_1 - q_1 \quad \dots \quad a_n - q_n] \quad (8)$$

$l_1(t) = 0$, $1 \leq m$ and $a_i - q_i$, $i = 1, 2, \dots, n$ are the estimated values. Notice that in the standard algorithm $\hat{y}(t) = c^T(t) \theta(t)$, i.e., $l_1(t) = 0$.

The adaptive algorithm is

$$\dot{\theta}(t) = \epsilon c(t) \theta(t) \quad \epsilon > 0 \quad (9)$$

We will now state and prove a theorem, on which our modified output error identification algorithm is based and justified.

Theorem 1 There exists a scalar $l^* > 0$ such that for all $l > l^*$ the modified output error algorithm described by (2)–(9) is asymptotically stable when the input $u(t)$ is bounded, i.e., all the signals in the system are bounded and $\lim_{t \rightarrow \infty} e(t) = 0$. If in addition the input $u(t)$ is persistently excited (PE) then $\lim_{t \rightarrow \infty} \theta(t) = \theta_0$ exponentially fast.

To prove Theorem 1, we need the following lemma which appears in [12].

Lemma 1 Let $Q(s)$ be a rational function in the complex variable s with real coefficients

$$Q(s) = \frac{\sum_{i=0}^m c_i s^i}{\sum_{i=0}^d d_i s^i} \quad c_i, d_i \neq 0$$

having a positive leading coefficient c_m/d_m , and satisfying

$$Q(s) \neq 0 \quad \text{in } \mathbb{R} \quad \text{Re}[s] \geq 0 \quad (10)$$

Also, if $Q(s)$ vanishes at infinity, the multiplicity of the zero is not greater than 1, i.e., $m \geq n - 1$.

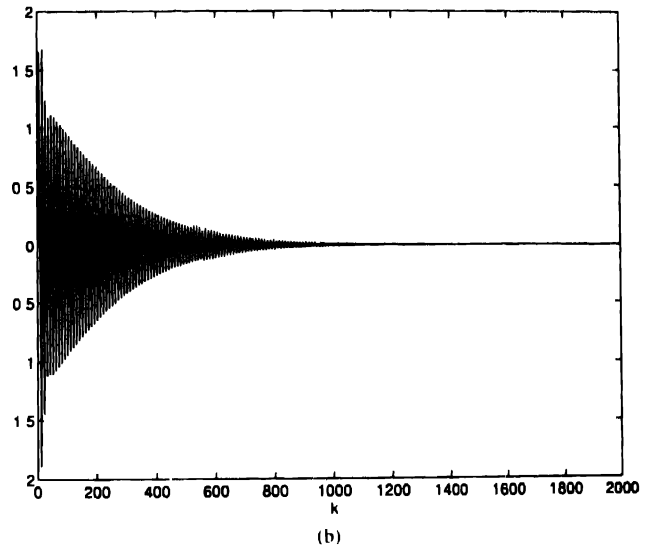
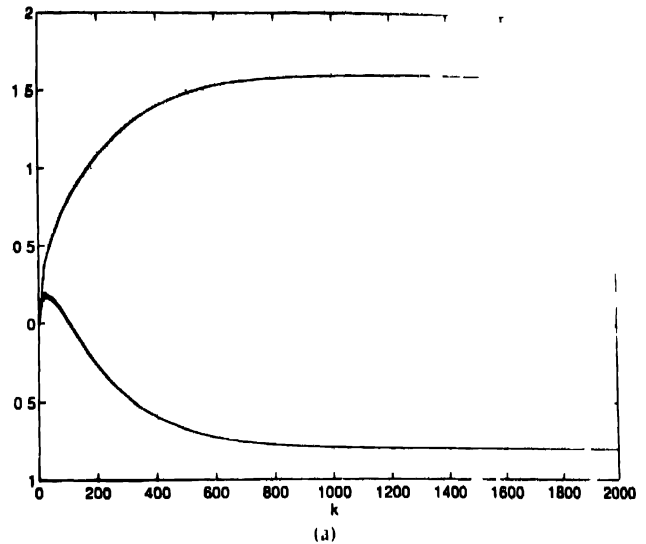


Fig. 1 (a) Estimated parameters versus sampling instants—modified algorithm, discrete time case. (b) Output error versus sampling instants—modified algorithm, discrete time case.

Let l be a positive finite number which is sufficiently large. Then the rational function $F(s)$ defined by

$$F(s) = \frac{Q(s)}{1 + lQ(s)} \quad (11)$$

is strictly positive real (SPR).

Proof of Theorem 1 For the case $l = 0$ it was shown in [2] that

$$\dot{e}(t) = -H(s)(c^T \theta) \quad (12)$$

where

$$H(s) = \frac{q(s)}{a(s)} \quad (13)$$

and

$$\theta = \theta_0 - \theta \quad (14)$$

A similar derivation, which will be omitted for the sake of brevity, reveals that for the case of $\hat{y}(t)$ as in (6) with $l > 0$, (13) should be replaced by

$$H(s) = \frac{q(s)}{a(s) + lq(s)} \quad (15)$$

It is well known [1]–[2] that if $H(s)$ is an SPR function, then the system is asymptotically stable. If, in addition, the input $u(t)$ is PE, then the parameters converge exponentially to their correct values, i.e., $\lim_{t \rightarrow \infty} \theta = \theta_0$ exponentially fast.

Define now $Q(s) \triangleq q(s)/u(s)$ which satisfies the conditions of Lemma 1. Then, (15) becomes $H(s) = Q(s)/(1 + lQ(s))$ and according to Lemma 1 we can find $l > 0$ sufficiently large such that $H(s)$ is SPR. \square

III. MODIFIED ALGORITHM—DISCRETE TIME CASE

We consider a plant described by

$$y(k) = -\sum_1^m a_i y(k-i) + \sum_{j=0}^n b_j u(k-j) \quad (16)$$

$$= c_0^l (k-1)\theta_l \quad (17)$$

where

$$c_0^l(l-1) = [u(k) \quad u(k-m) \quad y(k-1) \quad y(k-n)] \quad (18)$$

$$\theta_l^T = [b_l \quad b \quad -a_1 \quad -a] \quad (19)$$

The plant is assumed to be stable, i.e., all the zeros of $a(z^{-1}) \triangleq 1 + \sum_{i=1}^m a_i z^{-i}$ are in the open unit disk. The output error is

$$e(k) = y(k) - y(k) \quad (20)$$

As in the continuous time case, the standard $y(l)$ is modified by using feedback in the system's realization.

$y(k)$ is defined as

$$y(k) = \frac{1}{1+l} [\alpha^l (k-1)\theta(k) + l y(k)] \quad l > 0 \quad (21)$$

where

$$c_0^l(l-1) = [u(k) \quad u(k-m) \quad y(k-1) \quad y(k-n)] \quad (22)$$

$$\theta(l)^T = [b_0 \quad b \quad -a_1 \quad -a] \quad (23)$$

$b_i, i=0, 1, \dots, m$ and $a_i, i=1, 2, \dots, n$ are the estimated values. Notice that although (21) is algebraically equivalent to

$$y(k) = \alpha^l (k-1)\theta(k) + l y(k) \quad (24)$$

which is the form we actually desire for our stability analysis, (24) is not directly implementable, whereas (21) is readily implementable. The adaptation algorithm is

$$\theta(k+1) = \theta(k) - \frac{e(k)}{1 + \frac{e(k)}{c_0^l(k)} c_0^l(k)} [c_0^l(k)^T \theta(k) - y(k+1)] \quad (25)$$

We state and prove Theorem 2 which is the discrete time analog of Theorem 1.

Theorem 2 There exists a scalar $l^* > 0$ such that for all $l > l^*$ the modified output error algorithm described by (16)–(25) is asymptotically stable when the input $u(l)$ is bounded, i.e., all the signals in the system are bounded and $\lim_{k \rightarrow \infty} e(k) = 0$. If, in addition, the input $u(l)$ is persistently excited (PE), then $\lim_{k \rightarrow \infty} \theta(k) = \theta_0$ exponentially fast.

To prove Theorem 2 we need the following lemma which appears in [13].

Lemma 2 The discrete transfer function

$$Q(z^{-1}) = \frac{1}{1 + \sum_{i=1}^n a_i z^{-i}} \quad (26)$$

is SPR if

$$\sum_i |a_i| < 1 \quad (27)$$

Proof of Theorem 2 A similar derivation to the one carried out in [2], for the case $l = 0$, yields for the proposed case of $y(k)$, as in (21) with $l > 0$ the following expression

$$e(k) = -H(z^{-1})(\alpha^l (k-1)\theta(k)) \quad (28)$$

$$H(z^{-1}) = \frac{\frac{1}{1+l}}{1 + \frac{1}{1+l} \sum_{i=1}^n a_i z^{-i}} \quad (29)$$

where

$$\theta(k) = \theta(k) - \theta_0 \quad (30)$$

It is well known [1]–[3] that if $H(z^{-1})$ is an SPR function, then the system is asymptotically stable. If, in addition, the input $u(k)$ is PE, then the parameters converge exponentially to their correct values, i.e., $\lim_{k \rightarrow \infty} \theta(k) = \theta_0$ exponentially fast.

Obviously, there exists a sufficiently large l such that $1/(1+l) \sum_{i=1}^n |a_i| < 1$ and according to Lemma 2, $H(z^{-1})$ in (29) is SPR for such an l . \square

IV. THE DESIGN PROCEDURE

We assume now that we have some *a priori* knowledge about the denominator of the plant. We will point out how this knowledge allows us to design the feedback gain l which eliminates the SPR requirement.

In the continuous time case, we assume that $a(s)$, the denominator of the plant, belongs to the set of interval stable polynomials \mathbf{K} , i.e.,

$$a(s) = s^n + \sum_{i=1}^n a_i s^{n-i} \quad (31)$$

$$a_i \in [\underline{a}_i, \bar{a}_i] \quad i=1, 2, \dots, n \quad (32)$$

and $\underline{a}_i, \bar{a}_i, i=1, 2, \dots, n$ are given real numbers satisfying $\underline{a}_i < \bar{a}_i, i=1, 2, \dots, n$. The first step of the design procedure is to choose an arbitrary stable polynomial $q(s)$ of degree n . As discussed in [12], if

$$l > - \min_{\mathbf{K}} \left\{ \operatorname{Re} \left[\frac{a(j\omega)}{q(j\omega)} \right] \right\} \quad (33)$$

then the function $H(s)$ in (15) is SPR, and the adaptive system is asymptotically stable. Now let the four Kharitonov polynomials which are associated with the set \mathbf{K} [14] be denoted by $a_i(s), i=1, 2, 3, 4$. It has been proved in [15] that

$$\min_{\mathbf{K}} \left\{ \operatorname{Re} \left[\frac{a(j\omega)}{q(j\omega)} \right] \right\} = \min_{i=1, 2, 3, 4} \left\{ \operatorname{Re} \left[\frac{a_i(j\omega)}{q(j\omega)} \right] \right\} \quad (34)$$

Therefore, any feedback gain l which satisfies

$$l > - \min_{i=1, 2, 3, 4} \left\{ \operatorname{Re} \left[\frac{a_i(j\omega)}{q(j\omega)} \right] \right\} \quad (35)$$

guarantees the asymptotic stability of the adaptive system.

The right hand side of (35) can readily be computed, since it consists of four distinct fixed coefficients functions of a single variable ω .

In a more general case, where the denominator of the plant $a(s)$ belongs to a convex polytope of stable polynomials, it is evident that the coefficients can be bounded by a box, and the procedure carried out as in the interval case.

The design procedure in the discrete time case is carried out in a similar way and will be omitted for the sake of brevity.

Remark In [13], [16], and [17], a criterion was given for the asymptotic stability of the output error algorithm ([13] was generalized in [18]). This criterion states that for a large enough adaptation gain and/or input signal magnitude the system is asymptotically stable. In the present note, there are no restrictions on the adaptation gain or the input signal magnitude.

V SIMULATION RESULTS

In the discrete-time case, we investigate the example in [19], note that it is only the example which is being used and not the technique proposed in [19]. The boundedness conjecture in [19] was disproved in [20]. The plant is a second-order one with two unknown parameters, a_1, a described by

$$\begin{aligned} y(l) &= -a_1 y(k-1) - a y(k-2) + u(l) \\ &= 1.6y(k-1) - 0.8y(k-2) + u(k) \end{aligned} \quad (36)$$

The true parameter vector is $\theta^T = [-a_1 \ -a] = [1.6 \ -0.8]$. The transfer function of the error system in the original algorithm is $H(z^{-1}) = 1/1 - 1.6z^{-1} + 0.8z^{-2}$ which is not an SPR transfer function. It can be shown that for $l > 1.4$ the modified $H(z^{-1})$ is SPR. For convenience we chose $l = 3$. Now $H(z^{-1}) = 0.25/1 - 0.1z^{-1} + 0.2z^{-2}$ is an SPR transfer function.

In [19] it was found (by trial and error) that if the adaptation gain $\epsilon = 0.001773$ then the output error algorithm is not stable. We simulate this example with $\epsilon = 0.001$. Fig. 1(a) contains the plots of the estimated parameters versus time $\theta(l)$. Fig. 1(b) contains the plot of the output error versus time $e(l)$. This figure shows that, contrary to the original algorithm, in the modified algorithm the parameters converge to their correct values and the output error approaches zero.

VI CONCLUSION

The SPR condition is a major drawback of the output error algorithm because the denominator of the SPR transfer function has to be the denominator of the unknown plant. In the discrete time case it is possible to overcome this condition while in the continuous time case only sufficient conditions were found to satisfy the SPR condition. In the present note it was shown that by a slight modification of the output error algorithm we eliminate the SPR condition both in the continuous and discrete time cases. An *a priori* knowledge about the denominator of the plant allows us to design a feedback gain which guarantees the asymptotic stability of the algorithm. The design algorithm is very simple and by simulation we demonstrate the advantages of the modified algorithm over the original one.

REFERENCES

- [1] Y. D. Landau, *The Model Reference Approach*, New York: Marcel Dekker, 1976.
- [2] B. D. O. Anderson, R. R. Bitmead, C. R. Johnson Jr., P. V. Kokotovic, R. I. Kosut, I. M. Y. Mareels, L. Praly, and B. D. Riedels, *Stability of Adaptive Systems: Passivity and Averaging Analysis*, Cambridge, MA: MIT Press, 1986.
- [3] G. C. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction and Control*, Englewood Cliffs, NJ: Prentice Hall, 1984.
- [4] S. Dasgupta and A. Bhagwat, "Conditions for designing strictly positive real transfer functions for adaptive output error identification," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 731-736, July 1987.
- [5] S. Dasgupta, "A Kharitonov like theorem for systems under nonlinear passive feedback," in *Proc. 26th CDC*, Dec. 1987, pp. 2062-2063.
- [6] N. K. Bose and J. E. Delansky, "Boundedness of strictly positive real rational functions," *IEEE Trans. Circuits Syst.*, vol. CAS-36, pp. 103-104, Mar. 1989.
- [7] A. Katbab and F. I. Jury, "On the strictly positive real polynomials," *IEEE Trans. Automat. Contr.*, vol. AC-35, pp. 1031-1032, Dec. 1990.
- [8] Y. Q. Shi, "Robust (strictly) positive rational functions," *Circuit Syst.*, vol. 39, pp. 552-554, May 1991.
- [9] B. D. O. Anderson, S. Dasgupta, P. Khargonekar, I. J. K. Pappas, and M. Mansour, "Robust strict positive realness characterization of a class of transfer functions," *IEEE Trans. Circuits Syst.*, vol. 37, pp. 869-876, July 1990.
- [10] C. V. Hollot, L. Huang, and Z. L. Xu, "Designing strictly positive real transfer function families: A necessary and sufficient condition for low degree and structured families," in *Proc. MINS*, 1990.
- [11] A. Betser and E. Zehavi, "Design of robust strictly positive real transfer functions," *IEEE Trans. Circuits Syst.*, vol. 40, pp. 573-580, Sep. 1993.
- [12] E. Zehavi, "A sufficient condition for output feedback stabilization of uncertain systems," *IEEE Trans. Automat. Contr.*, vol. AC-33, pp. 1055-1057, Nov. 1986.
- [13] M. Tomizuka, "Parallel MRAS without compensation block," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 505-506, Apr. 1982.
- [14] V. I. Kharitonov, "Asymptotic stability of an equilibrium position of a family of systems of linear differential equations," *Differential Uravnen*, vol. 14, pp. 2086-2088, 1978.
- [15] H. Chapellat, M. Dahleh, and S. P. Bhattacharyya, "On robust nonlinear stability of interval control systems," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 59-67, Jan. 1991.
- [16] B. A. Altay, "Elimination of real positivity and error filtering in parallel MRAS," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 1017-1019, Nov. 1984.
- [17] M. Tomizuka, "On relaxation of SPR condition in parallel MRAS continuous time case," *IEEE Trans. Automat. Contr.*, vol. AC-33, pp. 976-979, Oct. 1988.
- [18] D. A. Lawrence and C. R. Johnson Jr., "Recursive parameter identification algorithm stability analysis via pi sharing," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 16-24, Jan. 1986.
- [19] D. A. Schoenwald and P. V. Kokotovic, "Boundedness conjecture for an output error adaptive algorithm," *Int. J. Adaptive Contr. Signal Processing*, vol. 4, pp. 27-47, 1990.
- [20] D. A. Lawrence, W. A. Schares, and W. Ren, "Parameter drift instability in disturbed free adaptive systems," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 584-587, Apr. 1993.

Comments on "Explicit Asymmetric Bounds for Robust Stability of Continuous and Discrete-Time Systems" and

Jian Xiao

Abstract—This note comments the results of a recent paper,¹ we point out that Theorem 2 is incorrectly stated. A correct version of the theorem is provided.

In the recent note,¹ the authors gave sufficient conditions for the robust stability of systems with parameter uncertainty. Since both the signs and the ranges of the uncertain parameters are taken into consideration, the proposed criterions are less conservative than the previous results.

In this note, we identify a minor error in their Theorem 2. In this theorem, the authors pointed out, for discrete time system

$$x(k+1) = (A + F)x(k) \quad (1)$$

with $F = \sum_{j=1}^m f_j F_j$, F is asymptotically stable if

$$\sum_k \lambda_k + \sum_j k f_j f_{j,j} < 1$$

where

$$\lambda_k = \begin{cases} \lambda_{\max}(P) & \text{for } k \geq 0 \\ \lambda_{\min}(P) & \text{for } k < 0 \end{cases} \quad (2)$$

$$f_{j,j} = \begin{cases} \lambda_{\max}(F_j) & \text{for } k, k_j \geq 0 \\ \lambda_{\min}(F_j) & \text{for } k, k_j < 0 \end{cases}$$

Manuscript received September 6, 1994.

The author is with the Department of Electrical Engineering, Southwest Jiaotong University, Chengdu 610031, Sichuan, P. R. China.

IEEE Log Number 9407005.

¹Z. Gao and P. J. Antsaklis, *IEEE Trans. Automat. Contr.*, vol. 38, pp. 332–335, 1993.

$$P = (E_i^T P A + A^T P E_i)/2, \quad i = 1, \dots, m$$

$$I_{i,j} = F_i^T P E_j/2, \quad i, j = 1, \dots, m$$

P is the solution of equation

$$A^T P A - P + 2I = 0$$

Since $F_{i,j}$ is generally not a Hermitian matrix, Lemma 2¹ is not valid here, and the theorem is generally not true. To give a correct statement, we first define two kind of matrices:

$$F_{i,j} = (F_i^T P F_j + L_i^T P E_j)/2, \quad i, j = 1, \dots, m, \quad i < j$$

$$G_i = F_i^T P E_i/2, \quad i = 1, \dots, m$$

note that both $F_{i,j}$ and G_i are Hermitian matrices, and

$$\sum_i G_i + \sum_{i,j} F_{i,j} = \sum_j F_j$$

We can give the correct version of this theorem as follows:

Theorem 1 The system in (1) is asymptotically stable if

$$\sum_k \lambda_k + \sum_i k q_i + \sum_{i,j} k, k_j f_{i,j} < 1$$

where λ_k is defined by (2) and

$$q_i = \lambda_{\max}(G_i), \quad i = 1, 2, \dots, m$$

$$f_{i,j} = \begin{cases} \lambda_{\max}(F_{i,j}) & \text{for } k, k_j \geq 0 \\ \lambda_{\min}(F_{i,j}) & \text{for } k, k_j < 0 \end{cases}, \quad i, j = 1, \dots, m, \quad i < j$$

Scanning the Issue*

Adaptive Control of Plants with Unknown Hystereses, *Wang and Kokotovic*

This paper studies the problem of adaptively controlling systems with unknown hystereses. Hysteresis phenomena caused by stiction, magnetism or gears with backlash often severely limit system performance and give rise to inaccuracy or oscillations, even leading to instability. Since hysteresis characteristics are generally nondifferentiable nonlinearities, the existing results of adaptive control theory which deal primarily with linear or differentiable nonlinear systems are not applicable to systems with hysteresis. The authors develop a new method to parameterize the unknown variables of a hysteresis model and develop a hysteresis inverse. They present an adaptive method for a known linear plant following the hysteresis as well as for an unknown plant. An adaptive hysteresis inverse cascaded with the plant is employed to cancel the effects of hysteresis so that the remaining part of the controller can retain its linear structure. The analytical results prove the global boundedness of the closed loop signals for wide class of hysteresis models.

Robust Stability under A Class of Nonlinear Parametric Perturbations, *Fu, Dasgupta and Blondel*

In verifying robust stability in linear systems dependent on real valued uncertain parameters, there is a fundamental trade off between the generality of functional dependence on the unknown parameters and the complexity of the tests for verifying stability. Therefore a key engineering challenge in these problems is to identify structures of uncertainty that facilitate tractable analysis, while capturing practically important cases. This paper's contribution falls in this category. For a transfer function description of finite dimensional linear time invariant systems, the authors derive tractable necessary and sufficient conditions for robust stability when given a particular structure of nonlinear parameter dependence in the associated characteristic function. The authors motivate their structure as one that might commonly arise from a cascade of parameter dependent single input single output plants with a fixed controller. If each of these uncertain plants has a numerator and denominator that are powers of polynomials having affine dependence on independent parameters and the parameters themselves are restricted to a hyper rectangle in R^n , the structure considered in this paper results.

The necessary and sufficient conditions for robust Hurwitz stability on this structure are derived using the concept of the frequency dependent value set for the characteristic function of interest. In particular, the authors identify simple one dimensional subsets of the parameter set that must be tested to ensure robust stability over the entire parameter set. These subsets are the edges of the parameter set hyper rectangle and certain internal line segments that may be frequency dependent. Moreover, the paper shows special cases of this uncertainty for which these internal line segments remain independent of frequency. To facilitate construction of algorithms to test these geometric conditions, the authors go on to identify a complex-valued function of frequency with the property that its image avoids the negative real axis if and only if the original uncertain system is robustly Hurwitz stable.

Discrete-Time Observers for Singularly Perturbed Continuous-Time Systems, *Shouse and Taylor*

This paper deals with constructing observers for nonlinear continuous time singularly perturbed systems. The observer design concept adopted here is based on inversion of state to measurement maps. The two time scale nature of the systems is exploited by decomposing the problem into one of designing separate observers for the approximate slow and fast subsystem models. The resulting observers are of reduced order and are also implemented in a multirate fashion. As a result, significant savings in computational and/or memory requirements are obtained. Also, the stiffness inherent in a full order observer design is eliminated.

The development and analysis of stabilizing observer based control algorithms is not attempted in this paper. Nevertheless, the use of observer based feedback is utilized in a case study of velocity control of a two phase permanent magnet synchronous motor using only phase current measurements. It is demonstrated via simulations, that the reduced order observer designs proposed in this paper result in good performance.

Adaptive Back-Pressure Congestion Control Based on Local Information, *Tassiulas*

A new distributed control policy for a multiclass multistage queueing systems that models virtual circuit and datagram communication networks and a class of manufacturing systems is presented in this paper. The policy aims at controlling the congestion in the network by dynamically allocating the servers to the buffers, regulating the flows in the network, and routing the served traffic. The author proposes a policy called Adaptive Back Pressure (ABP) congestion policy that performs all these control functions in a distributed fashion based on local information. For arrival processes with bounded burstiness, the author shows that the ABP policy stabilizes the network in the sense that it ensures bounded backlog in the network as long as the utilization of each server is less than its service capacity. An adaptive version of the ABP policy is also discussed.

Stability of Queueing Networks and Scheduling Policies, *Kumar and Mevn*

The objective of this paper is to establish stability results for queueing networks that involve scheduling servers among several competing queues. Given that solutions to the stability problem (namely, by determining a steady state probability distribution) are rare outside the special class of queueing networks possessing a product form solution, the authors propose and develop a programmatic procedure for obtaining sufficient conditions for the stability of queueing networks operating under scheduling policies. This procedure uses linear programming to programmatically construct an appropriate quadratic Lyapunov function from which the stability results are deduced. This method can provide stability results that could be very difficult to obtain analytically, and it is applied very successfully to several example problems in the paper.

Second-order Properties of Families of Discrete-Event Systems, *Rajan and Agrawal*

This paper considers a class of timed discrete-event systems termed synchronous discrete event systems (SDFS). Any such system

*This section is written by the Transactions Editorial Board.

can be described directly in terms of a language/score space, as opposed to a state evolution mechanism. SDES provide a unified framework for modeling a large class of Petri nets and queueing systems including tandem, cycle, fork-join, and split-merge networks with general blocking and starvation. The objective of the paper is to study the concavity of the event counting process of SDES. In particular, the paper presents a sufficient condition on score spaces that ensures that the event counting process of one system dominates

the event counting processes of a collection of systems. This result may be applied to conflict-free Petri nets to deduce the concavity of the firing process as function of the initial marking or to the above-mentioned families of queueing networks to deduce concavity of the throughput as a function of various parameters such as buffer configuration, initial job configuration, and blocking parameters. Overall, the authors present a new and interesting perspective that covers and extends earlier results.

Editorial

The 1994 George S. Axelby Outstanding Paper Award

(Covering the Period January 1992–December 1993)

THE George S. Axelby Outstanding Paper Award is given each year by the Control Systems Society to the authors of outstanding papers published in the IEEE TRANSACTIONS ON AUTOMATIC CONTROL during the two calendar years preceding the year of the Award. At most three papers can be selected each year. Papers are judged on their originality, their potential impact on the foundations of control, their importance and practical significance in applications, and on their clarity.

Nominations for the Award were solicited this year through announcements in the IEEE CONTROL SYSTEMS MAGAZINE and the TRANSACTIONS ON AUTOMATIC CONTROL. The Axelby Award Committee makes use of selected reviews and panel evaluations to evaluate and rank those papers which are nominated. As a result of this process, the following paper was selected this year to receive this prestigious Award:

Sheng-Luen Chung, Stéphane Lafortune, and Feng Lin,
"Limited lookahead policies in supervisory control of
discrete-event systems," vol. 37, no. 12, pp. 1921–1935,
Dec. 1992.

Prof. Chung is with the Electrical Engineering Technology Department of the National Taiwan Institute of Technology, Taipei, Republic of China. Prof. Lafortune is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, 48109 USA. Prof. Lin is with the Department of Electrical and Computer Engineering, Wayne State University, Detroit, Michigan, 48202 USA.

The authors received their Award during the Luncheon Award Ceremonies of the 33rd IEEE Conference on Decision and Control, at Lake Buena Vista, Florida, USA on December 15, 1994. The Award consists of a certificate and plaque and is named for George S. Axelby, founding editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL.

The following "Scanning the Issue" paragraph, written by Associate Editor C. G. Cassandras, is reprinted here again as a service to the reader:

One approach toward controlling discrete-event systems involves the design of "supervisors" whose purpose it is, for example, to ensure that the system reaches certain desirable state sets or avoids undesirable ones.

The design of supervisors is based on finite automata models both for the system to be controlled and for the process that describes its desired behavior. For complex systems, however, this task becomes extremely difficult for a number of reasons; for instance, one cannot overcome the exponential growth of the state space that accompanies complexity. The authors introduce an on-line scheme based on the idea of projecting the behavior of the system for N steps into the future after every event occurrence and then selecting a control action based on this limited information. This gives rise to various "limited lookahead policies" (LLP) resulting in an LLP supervisor. The authors also introduce the notion of an "attitude" regarding the control action selected, and they analyze a conservative and an optimistic attitude, for which they derive a number of properties. Lower bounds on the number of steps N defining the lookahead horizon are provided such that the LLP scheme's performance is at least as good as an off-line scheme with complete information. Finally, a simplified subway system is used as an example to illustrate the applicability of LLP supervisory control and related implementation issues.

Nominations for the 1995 George S. Axelby Outstanding Paper Award, covering the period January 1993–December 1994, are already open. Any reader who wishes to nominate a paper is urged to do so by the closing date of May 15, 1995. No nomination form is required. A letter identifying the paper by title, author, and issue of publication and including a description of the contribution of the paper should be submitted. Alternatively, an electronic message carrying the above information would be sufficient. Please send nominations, by May 15, 1995, to

Michael K. Sain
Freimann Professor of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556 USA
FAX: (219) 631-4393
email: jordan@medugorje.cc.nd.edu

MICHAEL K. SAIN
1994 Axelby Award Chairman

Abstract—For a system with hysteresis, we present a parameterized hysteresis model and develop a hysteresis inverse. We then design adaptive controllers with an adaptive hysteresis inverse for plants with unknown hystereses. A new adaptive controller structure is introduced which is capable of achieving a linear parameterization and a linear error model in the presence of a hysteresis nonlinearity. A robust adaptive law is used to update the controller parameters and hysteresis inverse parameters, which ensures the global boundedness of the closed-loop signals for a wide class of hysteresis models. Simulations show that the use of the adaptive hysteresis inverse leads to major improvements of system performance.

I. INTRODUCTION

HYSTERESIS phenomena caused by stiction, magnetism, or gears with backlash commonly exist in physical systems [1]–[6], and often severely limit system performance such as giving rise to undesirable inaccuracy or oscillations, even leading to instability. Hysteresis characteristics are generally nondifferentiable nonlinearities and usually unknown. Most results of adaptive control theory are for linear or differentiable nonlinear systems, and are not applicable to systems with nondifferentiable nonlinearities.

It is a challenging task of major practical interest to develop adaptive control schemes for systems with unknown hystereses. In this paper, we will pursue this task. The controlled plant consists of a linear part preceded by a hysteresis characteristic, that is, the hysteresis is present at the actuator of a linear part. The linear part can be either known or unknown, and the hysteresis is assumed to be unknown. The objective is to achieve stabilization and tracking in the presence of the unknown hysteresis. To solve this problem, we will use the “adaptive inverse” approach which was proposed in [7]–[9] for adaptive control of plants with nondifferentiable nonlinearities such as dead-zone and backlash. In [7], an adaptive dead-zone inverse control scheme was proposed for systems having an unknown dead zone at the input, using full state measurement. In [8], adaptive dead-zone inverse control designs for systems with unknown dead zones were developed, using only the output measurements. In [9], the adaptive inverse approach is used for adaptive control of systems were designed: one for systems with an unknown backlash and a known linear part, and the other for both the backlash and the linear

part unknown. In this paper, we will employ an adaptive hysteresis inverse cascaded with the plant to cancel the effects of hysteresis so that the remaining part of the controller can retain its linear structure.

The paper is organized as follows. In Section II, we present a hysteresis model characterized by a set of parameters, and formulate the control problem. In Section III, we present a mathematical model of a hysteresis inverse and its implementation. In Section IV, we develop a parameterization for an estimate of the hysteresis inverse. In Section V, we design an adaptive controller for plants with a known linear part and an unknown hysteresis. In Section VI, we present a solution to the more general adaptive control problem in which both the linear part and hysteresis are unknown. In Section VII, we present simulation results for the proposed adaptive control schemes.

Models of backlash, electromagnetic, and other types of hysteresis can be found in [1]–[6]. However, a general hysteresis model would not be convenient because of its complexity. We will use a simplified hysteresis model which captures most of the hysteresis characteristics and is useful for parameter adaptive control. We will show that our hysteresis model has a parameterizable right inverse which cancels the effect of the hysteresis when cascaded with the hysteresis. An adaptive hysteresis inverse is implemented with parameters updated on-line by adaptive laws.

When the adaptive hysteresis inverse is used for control, the effect of the hysteresis may not be completely cancelled, and there is a control error we express in two parts: first, a parameterizable part, and second, an unknown but bounded part.

For plants with an unknown linear part, the usual linear controller structure is modified in order to achieve a linear parameterization and a linear error model, which are both crucial for the development of robust adaptive laws. As a result of the linear parameterization of the adaptive hysteresis inverse and the new controller structure, robust adaptive laws can now be used to guarantee the closed-loop signal boundedness for a wide class of hysteresis characteristics. Our simulations show that major improvements of system performance have been achieved by the adaptive hysteresis inverse.

II. PLANT AND CONTROL OBJECTIVE

The plant to be controlled has a linear part and a hysteresis characteristic $H(\cdot)$ as its input:

$$y(t) = G(s)[u](t), \quad u(t) = H(v(t)) \quad (2.1)$$

where $G(s) = k_p(Z(s)/P(s))$, $Z(s)$, $P(s)$ are monic polynomials, k_p is a constant scalar, $y(t)$ is the measured output, $v(t)$ is the control input, and $H(\cdot) \triangleq H(m_t, c_t, m_b, c_b, m_r$.

Manuscript received March 2, 1992; revised December 2, 1993. Recommended by Past Associate Editor, A. Arapostathis. This work was supported by the National Science Foundation under Grants EEC-9203491 and ECS-9307545, by the Air Force Office of Scientific Research under Grant F-49620-92-J-0495, and by a Ford Motor Company grant.

G. Tao is with the Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903 USA.

P. V. Kokotović is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA.

IEEE Log Number 9407562

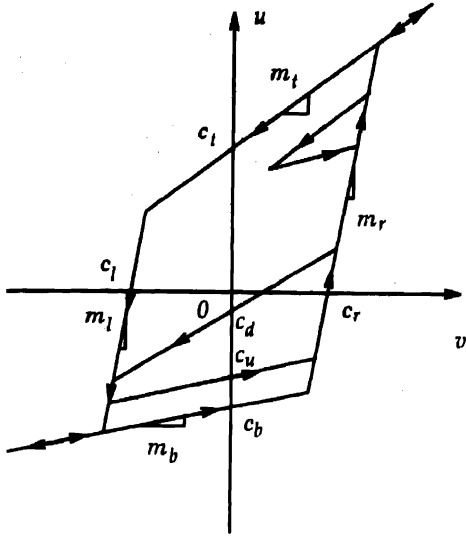


Fig. 1. Hysteresis model.

$c_r, m_l, c_l; \cdot$) is parameterized by constants $m_l, c_l, m_b, c_b, m_r, c_r, m_l, c_l$ and described by two half-lines (see Fig. 1):

$$u(t) = m_l v(t) + c_l, \quad v(t) > v_1 = \frac{c_l + m_l c_l}{m_l - m_t} \quad (2.2)$$

$$u(t) = m_b v(t) + c_b, \quad v(t) < v_2 = \frac{c_b + m_r c_r}{m_r - m_b} \quad (2.3)$$

and two line segments:

$$u(t) = m_r(v(t) - c_r), \quad v_2 \leq v(t) < v_3 = c_l + m_r c_r \quad (2.4)$$

$$u(t) = m_l(v(t) - c_l), \quad \frac{c_b + m_l c_l}{m_l - m_t} = v_4 < v(t) \leq v_1 \quad (2.5)$$

where v_1, v_2, v_3, v_4 are the values of $v(t)$ at the four opposite "corners" of the quadrilateral.

Along the segments, the time derivatives of $u(t), v(t)$ are of constant sign, namely, $\dot{u}(t) > 0, \dot{v}(t) > 0$ for $u(t) = m_r(v(t) - c_r)$, and $\dot{u}(t) < 0, \dot{v}(t) < 0$ for $u(t) = m_l(v(t) - c_l)$.

The hysteresis phenomena occur inside the loop formed by the half-lines (2.2)–(2.3) and the segments (2.4)–(2.5). Inside the hysteresis loop, the relationship between $u(t)$ and $v(t)$ is

$$u(t) = \begin{cases} m_t v(t) + c_d(t) & \text{for } \dot{v}(t) < 0 \\ m_b v(t) + c_u(t) & \text{for } \dot{v}(t) > 0 \end{cases} \quad (2.6)$$

where $c_d(t) \in (c_l, c_1), c_u(t) \in (c_2, c_b)$ are piecewise constant functions which depend on the point where $\dot{v}(t)$ changes its sign and on the past trajectories of $(v(t), u(t))$, with

$$c_1 = \begin{cases} (m_b - m_t) \frac{c_b + m_l c_l}{m_l - m_b} + c_b & \text{for } m_t < m_b \\ (m_b - m_t) \frac{c_b + m_r c_r}{m_r - m_b} + c_b & \text{for } m_t > m_b \\ c_b & \text{for } m_t = m_b \end{cases} \quad (2.7)$$

$$c_2 = \begin{cases} (m_t - m_b) \frac{c_l + m_r c_r}{m_r - m_t} + c_l, & \text{for } m_t > m_b \\ (m_t - m_b) \frac{c_l + m_l c_l}{m_l - m_t} + c_l, & \text{for } m_t < m_b \\ c_l & \text{for } m_t = m_b. \end{cases} \quad (2.8)$$

The relationship (2.6) holds also for a part of one of the half-lines: when $m_t > m_b$, on the half-line (2.2) with

$v_1 < v(t) < v_3, u(t) = m_t v(t) + c_l$ for $\dot{v}(t) < 0$; when $m_t < m_b$, on the half-line (2.3) with $v_4 < v(t) < v_2, u(t) = m_b v(t) + c_b$ for $\dot{v}(t) > 0$.

The signs of $\dot{u}(t)$ and $\dot{v}(t)$ are not restricted on other parts of these two half-lines: $u(t) = m_t v(t) + c_l, v(t) \geq v_3$; $u(t) = m_b v(t) + c_b, v(t) \leq v_4$; and $u(t) = m_t v(t) + c_l, v_1 < v(t) < v_3$ when $m_t < m_b$ or $u(t) = m_b v(t) + c_b, v_4 < v(t) < v_2$ when $m_t > m_b$. For example, the half-line $u(t) = m_b v(t) + c_b$ is bidirectional when $m_t > m_b$.

The motion of $u(t)$ and $v(t)$ on the half-lines (2.2)–(2.3) and the segments (2.4)–(2.5) and inside the hysteresis loop can be mathematically described as

$$\begin{aligned} \dot{u}(t) = \begin{cases} m_t \dot{v}(t) & \text{if } v(t) \geq v_3, \\ & \text{or if } v_4 < v(t) < v_3, \dot{v}(t) < 0, \\ & u(t) \neq m_l(v(t) - c_l) \text{ and} \\ & u(t) \neq m_b v(t) + c_b, \\ & \text{or if } v_4 < v(t) < v_3, \dot{v}(t) < 0, \\ & u(t) = m_b v(t) + c_b \text{ and } m_t < m_b \\ & \text{or if } v_4 < v(t) < v_3, \dot{v}(t) > 0, \\ & u(t) = m_t v(t) + c_l \text{ and } m_t < m_b \\ m_b \dot{v}(t) & \text{if } v(t) \leq v_4, \\ & \text{or if } v_4 < v(t) < v_3, \dot{v}(t) > 0, \\ & u(t) \neq m_r(v(t) - c_r) \text{ and} \\ & u(t) \neq m_t v(t) + c_l, \\ & \text{or if } v_4 < v(t) < v_3, \dot{v}(t) > 0, \\ & u(t) = m_t v(t) + c_l \text{ and } m_t > m_b \\ & \text{or if } v_4 < v(t) < v_3, \dot{v}(t) < 0, \\ & u(t) = m_b v(t) + c_b \text{ and } m_t > m_b \\ m_r \dot{v}(t) & \text{if } v_4 < v(t) < v_3, \dot{v}(t) > 0 \text{ and} \\ & u(t) = m_r(v(t) - c_r) \\ m_l \dot{v}(t) & \text{if } v_4 < v(t) < v_3, \dot{v}(t) < 0 \text{ and} \\ & u(t) = m_l(v(t) - c_l) \\ 0 & \text{if } \dot{v}(t) = 0. \end{cases} \quad (2.9) \end{aligned}$$

The model of the hysteresis and its two typical minor loops are shown in Fig. 1. This model captures most of the hysteresis phenomena described by more elaborate models in [5], e.g., it is an extended approximation of the ferromagnetic hysteresis model with $m_l = m_r = \infty, m_t = m_b = 0$ [6]. When $m_t = m_b$, the major hysteresis is the out-loop hysteresis, and the inner loops degenerate into line segments. When $m_t = m_b, m_r = m_l, c_l = -c_b > 0, c_r = -c_l > 0$, the hysteresis model (2.2)–(2.8) becomes the approximation model of [6] for the ferromagnetic hysteresis.

The assumptions about the hysteresis characteristic $H(\cdot)$ are

1) the hysteresis output $u(t)$ is not available for measurement;

2) $m_t > 0, m_b > 0, m_r > \max\{m_t, m_b\}, m_l > \max\{m_t, m_b\}, c_l \geq c_b, c_r \geq c_l$, and the values of these hysteresis parameters are unknown.

The assumptions for the linear part $G(s)$ of the plant are

3) $G(s)$ is minimum phase;

4) the relative degree n^* of $G(s)$ is known;

5) the degree n of $P(s)$ is known;

6) the sign of k_p is known.

The control objective is to design a feedback control $v(t)$ so that all closed-loop signals are bounded and the plant output

$y(t)$ tracks the output $y_m(t)$ of the reference model:

$$y_m(t) = W_m(s)[r](t) \quad (2.10)$$

where $W_m(s)$ is a stable rational transfer function of relative degree n^* , and $r(t)$ is a bounded piecewise continuous signal. Without loss of generality, we take $W_m(s) = P_m^{-1}(s)$, where $P_m(s)$ is a monic Hurwitz polynomial of degree n^* .

A linear model reference controller structure could be used to achieve the stated control objective of the hysteresis were absent, that is, $u(t) = v(t)$. In the presence of hysteresis, a linear controller alone cannot achieve the control objective. We propose to use an adaptive hysteresis inverse developed in the next sections to cancel the effect of hysteresis.

III. HYSTERESIS INVERSE

Let $u_d(t)$ be a control signal to be designed. As the inverse of the hysteresis (2.2)–(2.6), we use the hysteresis-like characteristic $HI(\cdot)$:

$$\begin{aligned} v(t) &= HI(u_d(t)) \\ &\triangleq HI(m_t, c_t, m_b, c_b, m_l, c_l; u_d(t)) \end{aligned} \quad (3.1)$$

which is defined by the half-lines (see Fig. 2):

$$v(t) = \frac{1}{m_t}(u_d(t) - c_t), \quad u_d(t) > u_1 = \frac{m_l(m_t c_l + c_t)}{m_l - m_t} \quad (3.2)$$

$$v(t) = \frac{1}{m_b}(u_d(t) - c_b), \quad u_d(t) < u_2 = \frac{m_l(m_b c_l + c_b)}{m_l - m_b} \quad (3.3)$$

and the line segments:

$$v(t) = \frac{1}{m_t} u_d(t) + c_t, \quad u_2 \leq u_d(t) \leq u_3 = \frac{m_l(m_t c_l + c_t)}{m_l - m_t} \quad (3.4)$$

$$v(t) = \frac{1}{m_l} u_d(t) + c_l, \quad \frac{m_l(m_b c_l + c_b)}{m_l - m_b} = u_4 < u_d(t) \leq u_1. \quad (3.5)$$

Along the segments, the time derivatives of $v(t)$, $u_d(t)$ are of constant sign, namely, $\dot{u}_d(t) > 0$, $\dot{v}(t) > 0$ for $v(t) = (1/m_t)u_d(t) + c_t$, and $\dot{u}_d(t) < 0$, $\dot{v}(t) < 0$ for $v(t) = (1/m_l)u_d(t) + c_l$.

Inside the loop formed by the half-lines (3.2)–(3.3) and the segments (3.4)–(3.5), the output of the hysteresis inverse $v(t)$ is defined as

$$v(t) = \begin{cases} \frac{1}{m_t}(u_d(t) - \hat{c}_d(t)) & \text{for } \dot{u}_d(t) < 0 \\ \frac{1}{m_b}(u_d(t) - \hat{c}_u(t)) & \text{for } \dot{u}_d(t) > 0 \end{cases} \quad (3.6)$$

where $\hat{c}_d(t) \in (c_t, c_1)$, $\hat{c}_u(t) \in (c_2, c_b)$ are piecewise constant functions which depend on the point where $\dot{u}_d(t)$ changes its sign and on the past trajectories of $(v(t), u_d(t))$, with c_1 and c_2 defined in (2.7) and (2.8).

The relationship (3.6) holds also for a part of one of the half-lines of the hysteresis inverse $HI(\cdot)$: when $m_t > m_b$, on the common part of the half-line (3.2) and the hysteresis loop, $v(t) = (1/m_t)(u_d(t) - c_t)$ for $\dot{u}_d(t) < 0$; when $m_t < m_b$, on the common part of the half-line (3.3) and the hysteresis loop, $v(t) = (1/m_b)(u_d(t) - c_b)$ for $\dot{u}_d(t) > 0$. $u_d(t)$ and $\dot{v}(t)$

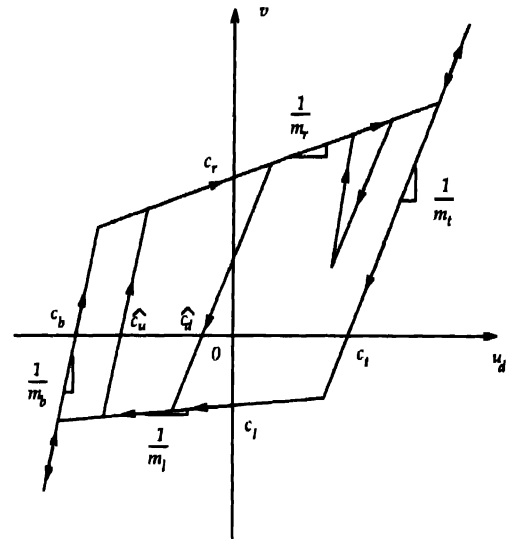


Fig. 2 Hysteresis inverse

are allowed to have both positive and negative signs on other parts of these two half-lines.

The model of the hysteresis inverse and its two typical minor loops are shown in Fig. 2.

The motion of $u_d(t)$ and $v(t)$ on the half-lines (3.2)–(3.3) and the segments (3.4)–(3.5) and inside the loop is mathematically described as

$$\dot{v}(t) = \begin{cases} \frac{1}{m_t} \dot{u}_d(t) & \text{if } u_d(t) \geq u_3, \\ & \text{or if } u_4 < u_d(t) < u_3, u_d(t) < 0, \\ & v(t) \neq \frac{1}{m_t} u_d(t) + c_t \text{ and} \\ & v(t) \neq \frac{1}{m_l} (u_d(t) - c_b), \\ & \text{or if } u_4 < u_d(t) < u_3, u_d(t) < 0, \\ & v(t) = \frac{1}{m_t} (u_d(t) - c_b) \text{ and } m_t < m_b \\ & \text{or if } u_4 < u_d(t) < u_3, u_d(t) > 0, \\ & v(t) = \frac{1}{m_l} (u_d(t) - c_t) \text{ and } m_l < m_b \\ \frac{1}{m_b} \dot{u}_d(t) & \text{if } u_d(t) < u_4, \\ & \text{or if } u_4 < u_d(t) < u_3, u_d(t) > 0, \\ & v(t) \neq \frac{1}{m_t} u_d(t) + c_t \text{ and} \\ & v(t) \neq \frac{1}{m_l} (u_d(t) - c_t), \\ & \text{or if } u_4 < u_d(t) < u_3, \dot{u}_d(t) > 0, \\ & v(t) = \frac{1}{m_t} (u_d(t) - c_t) \text{ and } m_t > m_b \\ & \text{or if } u_4 < u_d(t) < u_3, \dot{u}_d(t) < 0, \\ & v(t) = \frac{1}{m_b} (u_d(t) - c_b) \text{ and } m_t > m_b \\ \frac{1}{m_t} \dot{u}_d(t) & \text{if } u_4 < u_d(t) < u_3, u_d(t) > 0 \text{ and} \\ & v(t) = \frac{1}{m_t} u_d(t) + c_t \\ \frac{1}{m_l} \dot{u}_d(t) & \text{if } u_4 < u_d(t) < u_3, \dot{u}_d(t) < 0 \text{ and} \\ & v(t) = \frac{1}{m_l} u_d(t) + c_l \\ 0 & \text{if } \dot{u}_d(t) = 0. \end{cases} \quad (3.7)$$

The proposed hysteresis inverse has the following properties:

Proposition 3.1 (Hysteresis Inverse): The characteristic $HI(\cdot)$ defined by (3.2)–(3.6) is the right inverse of the characteristic $H(\cdot)$ defined by (2.2)–(2.6) in the sense that

$$H(HI(u_d(t_0))) = u_d(t_0) \Rightarrow H(HI(u_d(t))) = u_d(t), \quad \forall t \geq t_0 \quad (3.8)$$

for any piecewise continuous $u_d(t)$ and any $t_0 \geq 0$.

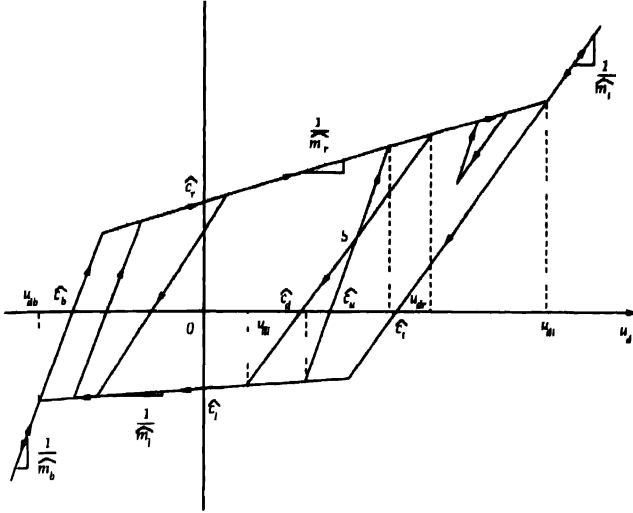


Fig. 3 Adaptive hysteresis inverse

The mapping (3.2)–(3.6) may fail to define a hysteresis inverse only if $u_d(t)$ is such that $v(t)$ and $u(t)$ never leave the hysteresis loop, that is, t_0 in (3.8) cannot be reached. Such a situation happens when $u_d(t)$ is inside the hysteresis inverse loop and $u(t) \neq u_d(t)$ initially and the motion of $u_d(t)$ is such that $u(t)$ never leaves the hysteresis loop to correct the error between $u_d(t)$ and $u(t)$. However, as $u_d(t)$ is the design signal at our disposal, an initialization of the hysteresis inverse by an appropriate choice of $u_d(t_0)$ can always make $v(t)$ and $u(t)$ leave the inside of the hysteresis loop at t_0 so that $u(t_0) = u_l(t_0)$ and then from Proposition 3.1 $u(t) = u_d(t)$ for any $t > t_0$.

In the adaptive control problem the hysteresis parameters are unknown so the exact hysteresis inverse (3.2)–(3.6) is not implementable. We propose to use an adaptive hysteresis inverse whose parameters are updated according to adaptive laws.

Let $\hat{m}_t, \hat{c}_t, \hat{m}_b, \hat{c}_b, \hat{m}_r, \hat{c}_r, \hat{m}_l, \hat{c}_l$ be the estimates of the unknown hysteresis parameters $m_t, c_t, m_b, c_b, m_r, c_r, m_l, c_l$. An estimated hysteresis inverse $\hat{HI}(\cdot)$ is defined as

$$v(t) = \hat{HI}(u_d(t)) = HI(\hat{m}_t, \hat{c}_t, \hat{m}_l, \hat{c}_b, \hat{m}_r, \hat{c}_r, \hat{m}_l, \hat{c}_l, u_d(t)) \quad (3.9)$$

Fig. 3 shows its characteristic with two typical minor loops, as well as two segments indicating two possible motion of $v(t)$ and $u_l(t)$ at point S . Now, $\hat{c}_u(t) \in (\hat{c}_r(t), \hat{c}_l(t))$, $\hat{c}_d(t) \in (\hat{c}_r(t), \hat{c}_l(t))$ depend on the point $(u_d(t), v(t))$ where $u_d(t)$ changes the sign, and on the past trajectories of the adaptive hysteresis inverse, where $\hat{c}_1(t)$ and $\hat{c}_2(t)$ are calculated from (2.5) and (2.6), but with the estimates of the hysteresis parameters.

For a fixed hysteresis inverse, the estimates $\hat{m}_t, \hat{c}_t, \hat{m}_b, \hat{c}_b, \hat{m}_r, \hat{c}_r, \hat{m}_l, \hat{c}_l$ are constants based on certain *a priori* knowledge of $m_t, c_t, m_b, c_b, m_r, c_r, m_l, c_l$. For an adaptive hysteresis inverse, $\hat{m}_t = \hat{m}_t(t)$, $\hat{c}_t = \hat{c}_t(t)$, $\hat{m}_b = \hat{m}_b(t)$, $\hat{c}_b = \hat{c}_b(t)$, $\hat{m}_r = \hat{m}_r(t)$, $\hat{c}_r = \hat{c}_r(t)$, $\hat{m}_l = \hat{m}_l(t)$, $\hat{c}_l = \hat{c}_l(t)$ are time-varying signals generated by an adaptive update law.

To ensure that (3.2)–(3.6) are implementable we require 7) positive constants $m_{t1}, m_{t2}, m_{b1}, m_{b2}, m_{r1}, m_{r2}, m_{l1}, m_{l2}$ and constants $c_{b1}, c_{b2}, c_{t1}, c_{t2}, c_{r1}, c_{r2}$ are known such that $m_{t1} \leq m_{t2}, m_{b1} \leq m_{b2}, m_{r1} \leq m_{r2}, m_{l1} \leq m_{l2}$, $\max\{m_{t2}, m_{l2}\} < m_{t0} \leq m_{t1}$ and $c_{t1} < c_{t2} < c_{r1} \leq c_{r2} < c_{b1} \leq c_{b2} < c_{l1} \leq c_{l2} < c_{d1} \leq c_{d2}$.

Even when not stated explicitly, we will use projection to ensure that the estimated hysteresis parameters satisfy the above inequalities.

A numerical scheme for implementing the estimated hysteresis inverse (3.8) is developed next.

At time $t = t_k$ we have the knowledge of $\hat{m}_t(t), \hat{c}_t(t), \hat{m}_b(t), \hat{c}_b(t), \hat{m}_r(t), \hat{c}_r(t), \hat{m}_l(t), \hat{c}_l(t), u_d(t), v(t)$ for $t = t_{k-1}$ and the knowledge of $u_d(t)$ for $t = t_k$. With the help of Fig. 3 where $S = (u_d(t_{k-1}), v(t_{k-1}))$ we introduce the quantities $u_{ll} \leq u_{dl} \leq u_{dl} \leq u_{dr}$

$$u_{dl} = \frac{\hat{m}_t(\hat{m}_l \hat{c}_l + \hat{c}_t)}{\hat{m}_l - \hat{m}_t} \quad (3.10)$$

$$\hat{c}_l = u_d(t_{k-1}) - \hat{m}_l v(t_{k-1}) \quad \hat{c}_u = u_l(t_{k-1}) - \hat{m}_l v(t_{k-1}) \quad (3.11)$$

$$\begin{aligned} & \frac{m_t(\hat{m}_l \hat{c}_l + \hat{c}_t)}{\hat{m}_l - \hat{m}_t} \quad \text{if } u_d(t_{k-1}) > u_l(t_k) \\ & \frac{m_t(\hat{m}_l \hat{c}_u + \hat{c}_t)}{\hat{m}_l - \hat{m}_t} \quad \text{if } u_d(t_{k-1}) < u_d(t_k) \end{aligned}$$

$$u_{dl} = \begin{cases} \frac{m_t(\hat{m}_l \hat{c}_l + \hat{c}_t)}{\hat{m}_l - \hat{m}_t} & \text{if } u_d(t_{k-1}) > u_d(t_k) \\ \frac{m_t(\hat{m}_l \hat{c}_u + \hat{c}_t)}{\hat{m}_l - \hat{m}_t} & \text{if } u_d(t_{k-1}) < u_d(t_k) \end{cases} \quad (3.12)$$

Then $v(t_k)$ is determined as

$$\begin{aligned} & v(t_{k-1}) \quad \text{if } u_d(t_k) = u_d(t_{k-1}) \\ & \frac{1}{m_t}(u_d(t_k) - \hat{c}_t) \quad \text{if } u_d(t_k) \geq u_{ll} \\ & \frac{m_t}{m_l}(u_d(t_k) - \hat{c}_l) \quad \text{if } u_d(t_k) < u_{ll} \\ & \frac{1}{m_t}u_d(t_k) + \hat{c}_r \quad \text{if } u_{ll} \geq u_d(t_k) \geq u_{dr} \\ & \frac{1}{m_t}u_d(t_k) + \hat{c}_l \quad \text{if } u_{dl} < u_l(t_k) \leq u_{ll} \\ & \frac{m_t}{m_l}(u_d(t_k) - \hat{c}_d) \quad \text{if } u_{dl} < u_l(t_k) < u_{dl} \text{ and } \\ & \quad \quad \quad u_d(t_{k-1}) > u_d(t_k) \\ & \frac{1}{m_t}(u_d(t_k) - \hat{c}_u) \quad \text{if } u_{dl} < u_d(t_k) < u_{dr} \text{ and } \\ & \quad \quad \quad u_l(t_{k-1}) < u_l(t_k) \end{aligned} \quad (3.13)$$

The implementation is dynamic because it uses not only $u_d(t_k)$ but also $u_d(t_{k-1})$ and $v(t_{k-1})$. It is important that this implementation does not need the knowledge of $u_d(t)$.

In our adaptive control schemes, the adaptive hysteresis inverse will be used in cascade with the plant with a hysteresis, and the signal $u_d(t)$ will be generated from a linear controller structure.

IV. PARAMETERIZATION

In the absence of the hysteresis, a linear controller generating $u(t)$ can be used to achieve the asymptotic tracking of $y_m(t)$ by $y(t)$. When the hysteresis is present, $v(t)$ is the accessible control which, by our design, is the output of the adaptive hysteresis inverse with input $u_d(t)$. A parameterization of the control error $u(t) - u_d(t)$ will help us to develop suitable

controller structures for generating $u_d(t)$ and adaptive laws for updating the estimates $\widehat{m}_t(t)$, $\widehat{c}_t(t)$, $\widehat{m}_b(t)$, $\widehat{c}_b(t)$, $\widehat{m}_r(t)$, $\widehat{c}_r(t)$, $\widehat{m}_l(t)$, $\widehat{c}_l(t)$ to implement an adaptive hysteresis inverse

To proceed, we define the following indicator functions

$$\chi_t(t) = \begin{cases} 1 & \text{if } (v(t), u(t)) \in \{(v(t), u(t)) \mid u(t) \\ & = m_t v(t) + c_t\} \\ 0 & \text{otherwise} \end{cases} \quad (4.1)$$

$$\chi_b(t) = \begin{cases} 1 & \text{if } (v(t), u(t)) \in \{(v(t), u(t)) \mid u(t) \\ & = m_b v(t) + c_b\} \\ 0 & \text{otherwise} \end{cases} \quad (4.2)$$

$$\chi_r(t) = \begin{cases} 1 & \text{if } (v(t), u(t)) \in \{(v(t), u(t)) \mid u(t) \\ & = m_r(v(t) - c_r)\} \\ 0 & \text{otherwise} \end{cases} \quad (4.3)$$

$$\chi_l(t) = \begin{cases} 1 & \text{if } (v(t), u(t)) \in \{(v(t), u(t)) \mid u(t) \\ & = m_l(v(t) - c_l)\} \\ 0 & \text{otherwise} \end{cases} \quad (4.4)$$

$$\chi_d(t) = \begin{cases} 1 & \text{if } (v(t), u(t)) \in \{(v(t), u(t)) \mid u(t) \\ & = m_t v(t) + c_d(t)\} \\ 0 & \text{otherwise} \end{cases} \quad (4.5)$$

$$\chi_u(t) = \begin{cases} 1 & \text{if } (v(t), u(t)) \in \{(v(t), u(t)) \mid u(t) \\ & = m_b v(t) + c_u(t)\} \\ 0 & \text{otherwise} \end{cases} \quad (4.6)$$

$$\widehat{\chi}_t(t) = \begin{cases} 1 & \text{if } (u_d(t), v(t)) \in \{(u_d(t), v(t)) \mid v(t) \\ & = \frac{1}{\widehat{m}_t(t)}(u_d(t) - \widehat{c}_t(t))\} \\ 0 & \text{otherwise} \end{cases} \quad (4.7)$$

$$\widehat{\chi}_b(t) = \begin{cases} 1 & \text{if } (u_d(t), v(t)) \in \{(u_d(t), v(t)) \mid v(t) \\ & = \frac{1}{\widehat{m}_b(t)}(u_d(t) - \widehat{c}_b(t))\} \\ 0 & \text{otherwise} \end{cases} \quad (4.8)$$

$$\widehat{\chi}_r(t) = \begin{cases} 1 & \text{if } (u_d(t), v(t)) \in \{(u_d(t), v(t)) \mid v(t) \\ & = \frac{1}{\widehat{m}_r(t)}(u_d(t) + \widehat{m}_r(t)\widehat{c}_r(t))\} \\ 0 & \text{otherwise} \end{cases} \quad (4.9)$$

$$\widehat{\chi}_l(t) = \begin{cases} 1 & \text{if } (u_d(t), v(t)) \in \{(u_d(t), v(t)) \mid v(t) \\ & = \frac{1}{\widehat{m}_l(t)}(u_d(t) + \widehat{m}_l(t)\widehat{c}_l(t))\} \\ 0 & \text{otherwise} \end{cases} \quad (4.10)$$

$$\widehat{\chi}_d(t) = \begin{cases} 1 & \text{if } (u_d(t), v(t)) \in \{(u_d(t), v(t)) \mid v(t) \\ & = \frac{1}{\widehat{m}_t(t)}(u_d(t) - \widehat{c}_d(t)), \\ & \widehat{c}_b(t) < \widehat{c}_d(t) < \widehat{c}_t(t)\} \\ 0 & \text{otherwise} \end{cases} \quad (4.11)$$

$$\widehat{\chi}_u(t) = \begin{cases} 1 & \text{if } (u_d(t), v(t)) \in \{(u_d(t), v(t)) \mid v(t) \\ & = \frac{1}{\widehat{m}_b(t)}(u_d(t) - \widehat{c}_u(t)), \\ & \widehat{c}_b(t) < \widehat{c}_u(t) < \widehat{c}_t(t)\} \\ 0 & \text{otherwise} \end{cases} \quad (4.12)$$

In defining these indicator functions, we do not repeatedly count any intersection of the half-lines and the segments, e.g., at the top right corner of $\widehat{HI}(\cdot)$ we define $\widehat{\chi}_l(t) = 1$ and $\widehat{\chi}_r(t) = 0$, and when it goes down passing the $\widehat{c}_b(t)$ point, we define $\widehat{\chi}_b(t) = 1$ and $\widehat{\chi}_d(t) = 0$. With this constraint, only one of these functions is nonzero at any given time t .

$$\widehat{\chi}_t(t) + \widehat{\chi}_b(t) + \widehat{\chi}_r(t) + \widehat{\chi}_l(t) + \widehat{\chi}_d(t) + \widehat{\chi}_u(t) = 1 \quad (4.13)$$

$$\widehat{\chi}_k^2(t) = \widehat{\chi}_k(t) \quad \widehat{\chi}_i(t)\widehat{\chi}_j(t) = 0 \quad i \neq j, i, j \in \{t, b, r, l, d, u\} \quad (4.14)$$

All the equalities in (4.1)–(4.12) are understood in the sense of (2.2)–(2.5) and the adaptive version of (3.2)–(3.5), that is, inequalities apply whenever they are feasible. We note that the condition 7) is necessary to ensure a well defined adaptive hysteresis inverse with the properties (4.13) and (4.14).

Using (2.2)–(2.6), (4.1)–(4.6), (4.13), we express $u(t)$ as

$$\begin{aligned} u(t) &= \chi_t(t)(m_t v(t) + c_t) + \chi_b(t)(m_b v(t) + c_b) \\ &\quad + \chi_r(t)(m_r(v(t) - c_r)) + \chi_l(t)(m_l(v(t) - c_l)) \\ &\quad + \chi_d(t)(m_t v(t) + c_d(t)) + \chi_u(t)(m_b v(t) + c_u(t)) \\ &= u_d(t) + \widehat{\chi}_t(t)(m_t v(t) + c_t - \widehat{\chi}_t(t)u_d(t)) \\ &\quad + \widehat{\chi}_b(t)(m_b v(t) + c_b - \widehat{\chi}_b(t)u_d(t)) \\ &\quad + \widehat{\chi}_r(t)(m_r(v(t) - c_r) - \widehat{\chi}_r(t)u_d(t)) \\ &\quad + \widehat{\chi}_l(t)(m_l(v(t) - c_l) - \widehat{\chi}_l(t)u_d(t)) \\ &\quad + \widehat{\chi}_d(t)(m_t v(t) + c_d(t) - \widehat{\chi}_d(t)u_d(t)) \\ &\quad + \widehat{\chi}_u(t)(m_b v(t) + c_u(t) - \widehat{\chi}_u(t)u_d(t)) + d_1(t) \end{aligned} \quad (4.15)$$

where

$$\begin{aligned} d_1(t) &= (\chi_t(t) - \widehat{\chi}_t(t))(m_t v(t) + c_t) \\ &\quad + (\chi_b(t) - \widehat{\chi}_b(t))(m_b v(t) + c_b) \\ &\quad + (\chi_r(t) - \widehat{\chi}_r(t))(m_r(v(t) - c_r)) \\ &\quad + (\chi_l(t) - \widehat{\chi}_l(t))(m_l(v(t) - c_l)) \\ &\quad + (\chi_d(t) - \widehat{\chi}_d(t))(m_t v(t) + c_d(t)) \\ &\quad + (\chi_u(t) - \widehat{\chi}_u(t))(m_b v(t) + c_u(t)) \end{aligned} \quad (4.16)$$

Since a projection based on the condition 7) ensures that all estimated hysteresis parameters are bounded, the adaptive hysteresis inverse loop is bounded. If $v(t)$ is large such that $(v(t), u_d(t))$ is outside the hysteresis inverse loop and $(u(t), v(t))$ is outside the hysteresis loop, then all $\chi(t)$'s and $\widehat{\chi}(t)$'s are zero except for $\chi_r(t)$ (or $\chi_l(t)$) and $\widehat{\chi}_r(t)$ (or $\widehat{\chi}_l(t)$), so that $d_1(t) = 0$ from (4.16). This implies that $d_1(t)$ is bounded whenever the estimated hysteresis parameters satisfy 7), which is crucial for our designs.

Using (3.2)–(3.6), (4.7)–(4.12), we express $v(t)$ as

$$\begin{aligned} v(t) = & \hat{\chi}_t(t) \left(\frac{1}{\widehat{m}_t(t)} (u_d(t) - \widehat{c}_t(t)) \right) \\ & + \widehat{\chi}_b(t) \left(\frac{1}{\widehat{m}_b(t)} (u_d(t) - \widehat{c}_b(t)) \right) \\ & + \widehat{\chi}_r(t) \frac{1}{\widehat{m}_r(t)} (u_d(t) + \widehat{m}_r(t) \widehat{c}_r(t)) \\ & + \widehat{\chi}_l(t) \left(\frac{1}{\widehat{m}_l(t)} (u_d(t) + \widehat{m}_l(t) \widehat{c}_l(t)) \right) \\ & + \widehat{\chi}_d(t) \left(\frac{1}{\widehat{m}_d(t)} (u_d(t) - \widehat{c}_d(t)) \right) \\ & + \widehat{\chi}_u(t) \left(\frac{1}{\widehat{m}_u(t)} (u_d(t) - \widehat{c}_u(t)) \right) \end{aligned} \quad (4.17)$$

In view of (4.14), from (4.17), we obtain

$$\widehat{\chi}_t(t) u_d(t) = \widehat{\chi}_t(t) (\widehat{m}_t(t) v(t) + \widehat{c}_t(t)) \quad (4.18)$$

$$\widehat{\chi}_b(t) u_d(t) = \widehat{\chi}_b(t) (\widehat{m}_b(t) v(t) + \widehat{c}_b(t)) \quad (4.19)$$

$$\widehat{\chi}_r(t) u_d(t) = \widehat{\chi}_r(t) (\widehat{m}_r(t) v(t) - \widehat{m}_r(t) \widehat{c}_r(t)) \quad (4.20)$$

$$\widehat{\chi}_l(t) u_d(t) = \widehat{\chi}_l(t) (\widehat{m}_l(t) v(t) - \widehat{m}_l(t) \widehat{c}_l(t)) \quad (4.21)$$

$$\widehat{\chi}_d(t) u_d(t) = \widehat{\chi}_d(t) (\widehat{m}_d(t) v(t) + \widehat{c}_d(t)) \quad (4.22)$$

$$\widehat{\chi}_u(t) u_d(t) = \widehat{\chi}_u(t) (\widehat{m}_u(t) v(t) + \widehat{c}_u(t)). \quad (4.23)$$

Introducing $\widehat{m}_r c_r(t) = \widehat{m}_r(t) \widehat{c}_r(t)$, $\widehat{m}_l c_l(t) = \widehat{m}_l(t) \widehat{c}_l(t)$, and

$$\theta_h(t) = (\widehat{m}_t(t), \widehat{c}_t(t), \widehat{m}_b(t), \widehat{c}_b(t), \widehat{m}_r(t), \widehat{m}_r c_r(t), \widehat{m}_l(t), \widehat{m}_l c_l(t))^T \quad (4.24)$$

$$\begin{aligned} \theta_h^* &= (m_t, c_t, m_b, c_b, m_r, m_r c_r, m_l, m_l c_l)^T, \\ \phi_h(t) &= \theta_h(t) - \theta_h^* \end{aligned} \quad (4.25)$$

$$\begin{aligned} \omega_h(t) = & (-\widehat{\chi}_t(t) + \widehat{\chi}_d(t)) v(t), -\widehat{\chi}_t(t), \\ & -(\widehat{\chi}_b(t) + \widehat{\chi}_u(t)) v(t), -\widehat{\chi}_b(t), -\widehat{\chi}_r(t) v(t), \\ & \widehat{\chi}_r(t), -\widehat{\chi}_l(t) v(t), \widehat{\chi}_l(t))^T \end{aligned} \quad (4.26)$$

from (4.15), (4.18)–(4.26), we have

$$u(t) = u_d(t) + \phi_h^T(t) \omega_h(t) + d_h(t) \quad (4.27)$$

where

$$\begin{aligned} d_h(t) = & d_1(t) + \widehat{\chi}_d(t) (c_d(t) - \widehat{c}_d(t)) \\ & + \widehat{\chi}_u(t) (c_u(t) - \widehat{c}_u(t)). \end{aligned} \quad (4.28)$$

Thus, we have expressed the control error $u(t) - u_d(t)$ as a sum of a parameterizable part and an unknown disturbance. The disturbance $d_h(t)$ has the following properties.

Proposition 4.1: The unparameterizable part $d_h(t)$ of the control error $u(t) - u_d(t)$ is bounded for any $t \geq 0$, and reduces to zero when the hysteresis parameter error $\phi_h(t) \rightarrow 0$.

The signals $\chi_t(t)$, $\chi_b(t)$, $\chi_r(t)$, $\chi_l(t)$, $\chi_d(t)$, and $\chi_u(t)$, which describe the motion of the hysteresis output $u(t) = H(v(t))$ [see the first expression for $u(t)$ in (4.15)], are not available for measurement. We choose to parameterize the control error $u(t) - u_d(t)$ in terms of the measured signals $\widehat{\chi}_t(t)$, $\widehat{\chi}_b(t)$, $\widehat{\chi}_r(t)$, $\widehat{\chi}_l(t)$, $\widehat{\chi}_d(t)$, and $\widehat{\chi}_u(t)$ [see the definition (4.26) for $\omega_h(t)$] which describe the motion of the adaptive hysteresis inverse output $v(t) = \widehat{H}(u_d(t))$ [see (4.17)].

The parameterization (4.27) can be made simpler for some special hysteresis characteristics.

Hysteresis with Equal Slopes $m_t = m_b$: The knowledge that $m_t = m_b$ can be used to choose $\widehat{m}_b(t) = \widehat{m}_t(t)$, $\widehat{c}_u(t) = \widehat{c}_d(t)$. In this case, we define

$$\theta_h^* = (m_t, c_t, c_b, m_r, m_r c_r, m_l, m_l c_l)^T \quad (4.29)$$

$$\begin{aligned} \theta_h(t) = & (\widehat{m}_t(t), \widehat{c}_t(t), \widehat{c}_b(t), \widehat{m}_r(t), \\ & \widehat{m}_r c_r(t), \widehat{m}_l(t), \widehat{m}_l c_l(t))^T \end{aligned} \quad (4.30)$$

$$\begin{aligned} \omega_h(t) = & (-\widehat{\chi}_t(t) + \widehat{\chi}_d(t) + \widehat{\chi}_b(t)) v(t), -\widehat{\chi}_t(t), -\widehat{\chi}_b(t), \\ & -\widehat{\chi}_r(t) v(t), \widehat{\chi}_r(t), -\widehat{\chi}_l(t) v(t), \widehat{\chi}_l(t))^T \end{aligned} \quad (4.31)$$

where

$$\begin{aligned} \widehat{\chi}_c(t) = & \begin{cases} 1 & \text{if } (u_d(t), v(t)) \in \{(u_d(t), v(t)) \mid v(t) \\ & = \frac{1}{\widehat{m}_r(t)} (u_d(t) - \widehat{c}_d(t)), \\ & \widehat{c}_b(t) < \widehat{c}_d(t) < \widehat{c}_t(t)\} \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (4.32)$$

Symmetric Hysteresis: When the hysteresis is symmetric, $m_t = m_b$, $m_r = m_l$, $c_t = -c_b > 0$, $c_r = -c_l > 0$, we define

$$\theta_h^* = (m_t, c_t, m_r, m_r c_r)^T \quad (4.33)$$

$$\theta_h(t) = (\widehat{m}_t(t), \widehat{c}_t(t), \widehat{m}_r(t), \widehat{m}_r c_r(t))^T \quad (4.34)$$

$$\begin{aligned} \omega_h(t) = & (-\widehat{\chi}_t(t) + \widehat{\chi}_d(t) + \widehat{\chi}_b(t)) v(t), -(\widehat{\chi}_t(t) - \widehat{\chi}_b(t)), \\ & -(\widehat{\chi}_r(t) + \widehat{\chi}_l(t)) v(t), \widehat{\chi}_r(t) - \widehat{\chi}_l(t))^T \end{aligned} \quad (4.35)$$

where $\widehat{\chi}_c(t)$ is also defined by (4.32), but with $\widehat{c}_b(t) = -\widehat{c}_t(t)$. In both cases, (4.27) and Proposition 4.1 hold.

V. ADAPTIVE HYSTERESIS INVERSE FOR $G(s)$ KNOWN

In this section, we design an adaptive control scheme for the plant (2.1) with a known linear part.

We use the following linear controller to generate $u_d(t)$:

$$u_d(t) = \theta_1^{*T} \omega_1(t) + \theta_2^{*T} \omega_2(t) + \theta_{20}^* y(t) + \theta_3^* r(t). \quad (5.1)$$

There are two designs for θ_1^* , θ_2^* , θ_{20}^* , and θ_3^* and $\omega_1(t)$ and $\omega_2(t)$.

I) For $a(s) = (1, s, \dots, s^{n-2})^T$, $\Lambda(s)$ being any monic Hurwitz polynomial of degree $n - 1$

$$\omega_1(t) = \frac{a(s)}{\Lambda(s)} [u_d](t), \quad \omega_2(t) = \frac{a(s)}{\Lambda(s)} [y](t) \quad (5.2)$$

$\theta_3^* = k_p^{-1}$, and $\theta_1^*, \theta_2^* \in R^{n-1}$, $\theta_{20}^* \in R$ satisfying the following Diophantine equation

$$\theta_1^{*T} a(s)P(s) + (\theta_2^{*T} a(s) + \theta_{20}^* \Lambda(s))k_p Z(s) = \Lambda(s)(P(s) - k_p \theta_3^* Z(s)P_m(s)). \quad (5.3)$$

II) For $a(s) = (1, s, \dots, s^{n-1})^T$, $\Lambda(s)$ being any monic Hurwitz polynomial of degree n , $\omega_1(t)$ and $\omega_2(t)$ still have the form (5.2), and $\theta_3^* = k_p^{-1}$, $\theta_1^*, \theta_2^* \in R^n$, $\theta_{20}^* = 0$ satisfying (5.3).

Then the signal $u_d(t)$ is applied to the adaptive hysteresis inverse (3.8) to generate $v(t)$ as the control input to the plant (2.1) which has a hysteresis at input.

To implement an adaptive hysteresis inverse, we can use the adaptive version of the continuous-time hysteresis inverse (3.7) or the discrete-time version (3.13). For the controller II, both the continuous-time and discrete-time adaptive hysteresis inverses can be implemented because, from (5.1) with $\theta_{20}^* = 0$, the derivative of $u_d(t)$, $\dot{u}_d(t)$, which is needed for implementing (3.7), is also available (assuming that $\dot{v}(t)$ is bounded). For the controller II, one can only use the discrete-time inverse (3.13) since, in this case, $u_d(t)$ depends on $\dot{y}(t)$, which is not measured. For the discrete-time inverse, there is no need to assume the boundedness of $\dot{v}(t)$.

The parameter matching equation (5.3) can be used to derive

$$u(t) = \theta_1^{*T} \frac{a(s)}{\Lambda(s)} [v](t) + \theta_2^{*T} \frac{a(s)}{\Lambda(s)} [y](t) + \theta_{20}^* y(t) + \theta_3^* W_m^{-1}(s) [y](t). \quad (5.4)$$

Using (2.9), (4.27), (5.1), (5.4), introducing $v(t) = y(t) - y_m(t)$, $F(s) = \theta_1^{*-1} W_m(s)(1 - \theta_3^{*T} (a(s)/\Lambda(s)))$, we obtain the tracking error equation

$$v(t) = F(s)[\phi_h^T \omega_h](t) + F(s)[d_h](t). \quad (5.5)$$

We note that $F(s)$ is a stable, strictly proper, and known transfer function.

Introducing

$$\xi_h(t) = \theta_h^T(t) \zeta_h(t) - F(s)[\theta_h^T \omega_h](t), \quad \zeta_h(t) = F(s)[\omega_h](t) \quad (5.6)$$

$$e_h(t) = v(t) + \xi_h(t) \quad (5.7)$$

we use the following adaptive law to update $\theta_h(t)$

$$\dot{\theta}_h(t) = \frac{\Gamma_h \zeta_h(t) e_h(t)}{1 + \zeta_h^T(t) \zeta_h(t) + \xi_h^2(t)} - \Gamma_h \sigma(\theta_h, M_h, \sigma_0) \theta_h(t) \quad (5.8)$$

where $\Gamma_h = \Gamma_h^T > 0$, and σ is a "switching-sigma" signal [10] using *a priori* knowledge of an upper bound M_h on the Euclidean norm $\|\theta_h^*\|$ of θ_h^* and a design parameter $\sigma_0 > 0$:

$$\sigma(\theta_h, M_h, \sigma_0) = \begin{cases} \text{if } \|\theta_h(t)\| < M_h \\ \left(\frac{\|\theta_h(t)\|}{M_h} - 1 \right) \\ \text{if } M_h \leq \|\theta_h(t)\| < 2M_h \\ \sigma_0 \\ \text{if } \|\theta_h(t)\| \geq 2M_h. \end{cases} \quad (5.9)$$

This adaptive law has the following properties

Lemma 5.1: The adaptive law (5.8) guarantees

- 1) $\theta_h(t)$, $\dot{\theta}_h(t)$, $(\epsilon_h^2(t)/1 + \zeta_h^T(t)\zeta_h(t) + \xi_h^2(t)) \in L_\infty$;
- 2) for some constants $k_1 > 0$, $k_2 > 0$, and all $t_2 > t_1 \geq 0$,

$$\int_{t_1}^{t_2} \|\dot{\theta}_h(t)\|^2 dt \leq k_1 + \int_{t_1}^{t_2} \frac{k_2}{1 + \zeta_h^T(t)\zeta_h(t) + \xi_h^2(t)} dt \quad (5.10)$$

$$\int_{t_1}^{t_2} \frac{\epsilon_h^2(t)}{1 + \zeta_h^T(t)\zeta_h(t) + \xi_h^2(t)} dt \leq k_1 + \int_{t_1}^{t_2} \frac{k_2}{1 + \zeta_h^T(t)\zeta_h(t) + \xi_h^2(t)} dt. \quad (5.11)$$

Proof With the substitution of (5.5) and (5.6) in (5.7), the estimation error $e_h(t)$ becomes

$$e_h(t) = \phi_h^T(t) \zeta_h(t) + d(t), \quad d(t) = F(s)[d_h](t). \quad (5.12)$$

Using (5.8) and (5.12), we express the time derivative of the positive-definite function $V_h(\phi_h) = \frac{1}{2} \phi_h^T \Gamma_h^{-1} \phi_h$ along the trajectories of (5.8), as

$$\begin{aligned} \dot{V}_h \leq & -\frac{\epsilon_h^2(t)}{2(1 + \zeta_h^T(t)\zeta_h(t) + \xi_h^2(t))} \\ & + \frac{d^2(t)}{2(1 + \zeta_h^T(t)\zeta_h(t) + \xi_h^2(t))} \\ & \sigma(\theta_h, M_h, \sigma_0) \phi_h^T(t) \theta_h(t). \end{aligned} \quad (5.13)$$

Since $d(t)$ is bounded, it follows from (5.13) that $\theta_h(t) \in L_\infty$, and $(\epsilon_h(t)/\sqrt{1 + \zeta_h^T(t)\zeta_h(t) + \xi_h^2(t)}), \theta_h(t) \in L_\infty$.

The definition of the signal $\sigma(t) = \sigma(\theta_h, M_h, \sigma_0)$ implies

$$\begin{aligned} \sigma^2(t) \|\theta_h(t)\|^2 \|\Gamma_h\|^2 & \leq k_0 \sigma(t) \|\theta_h(t)\| (\|\theta_h(t)\| - \|\theta_h^*\|) \\ & \leq k_0 \sigma(t) \phi_h^T(t) \theta_h(t) \end{aligned} \quad (5.14)$$

for some constant $k_0 > 0$. Then, using (5.13) and

$$\begin{aligned} \|\dot{\theta}_h(t)\|^2 & \leq 2 \frac{\|\Gamma_h\|^2 \|\zeta_h(t)\|^2 \epsilon_h^2(t)}{(1 + \zeta_h^T(t)\zeta_h(t) + \xi_h^2(t))^2} \\ & \quad + \|\Gamma_h\|^2 \sigma^2(\theta_h, M_h, \sigma_0) \|\theta_h(t)\|^2 \end{aligned} \quad (5.15)$$

we have part (2) of the lemma. \square

Lemma 5.1 only shows the boundedness of $\theta_h(t)$ and the L^2 properties (5.10)–(5.11) of the adaptive law (5.8). While the signal boundedness of the adaptive control system with a general hysteresis is still under investigation, we present the stability results for some special cases.

Theorem 5.1: When the hysteresis $H(\cdot)$ in (2.1) has two equal slopes, $m_t = m_b$ and (4.30), (4.31) are used for adaptation and implementation of the adaptive hysteresis inverse, all closed-loop signals are bounded.

The proof for Theorem 5.1 is given in Appendix A.

This boundedness result holds for the symmetric hysteresis case where (4.34), (4.35) are used for adaptation and implementation of the adaptive hysteresis inverse with a reduced system order.

VI. ADAPTIVE HYSTERESIS INVERSE FOR $G(s)$ UNKNOWN

The problem of adaptively controlling the plant (2.1) with an unknown hysteresis $H(\cdot)$ as well as an unknown linear part $G(s)$ challenges us with the task of estimating two sets of parameters: one from the hysteresis inverse, and the other from a linear controller structure. In this case, the linear controller structure (5.1) results in a nonlinear parameterization which is not suitable for parameter estimation.

Facing this difficulty, we need to modify the linear structure (5.1). Using (4.13), (4.18)–(4.23), (4.25), we express the hysteresis inverse (3.2)–(3.6) as

$$u_d(t) = -\theta_h^{*T} \omega_h(t) + \widehat{\chi}_d(t) \widehat{c}_d(t) + \widehat{\chi}_u(t) \widehat{c}_u(t). \quad (6.1)$$

Hence, the term $\theta_1^{*T} \omega_1(t)$ in (5.1) becomes

$$\begin{aligned} \theta_1^{*T} \omega_1(t) = & -\theta_1^{*T} \frac{a(s)}{\Lambda(s)} [\theta_h^{*T} \omega_h(t) \\ & + \theta_1^{*T} \frac{a(s)}{\Lambda(s)} [\widehat{\chi}_d \widehat{c}_d + \widehat{\chi}_u \widehat{c}_u](t)]. \end{aligned} \quad (6.2)$$

Introducing new regressors and parameters as

$$\omega_4(t) = \frac{A(s)}{\Lambda(s)} [\omega_h(t)], \quad \theta_4^* = -\theta_1^* \otimes \theta_h^* \quad (6.3)$$

$$\omega_5(t) = \frac{a(s)}{\Lambda(s)} [\widehat{\chi}_d \widehat{c}_d + \widehat{\chi}_u \widehat{c}_u](t), \quad \theta_5^* = \theta_1^* \quad (6.4)$$

where \otimes denotes the Kronecker product, $A(s) = (I_p, sI_p, \dots, s^{n-2}I_p)^T$, I_p is the $p \times p$ identity matrix with p being the dimension of θ_h^* , from (6.2), we obtain

$$\theta_1^{*T} \omega_1(t) = \theta_4^{*T} \omega_4(t) + \theta_5^{*T} \omega_5(t). \quad (6.5)$$

We should note that the dimension of θ_4^* and its estimate $\theta_4(t)$ is $p(n-1)$ (pn) for the controller I (controller II) of Section V. It is this enlarged parameterization that enables us to obtain a linear parameterization of the error system, which is suitable for controller adaptation. Physically, the hysteresis has eight different regions, so that the feedforward part $\theta_1^{*T}(t)\omega_1(t)$ has eight different forms. A natural parameterization turns out to be the one given in (6.5), with the parameter vector θ_4^* .

This new parameterization brings us to a new controller structure for adaptive control

$$\begin{aligned} u_d(t) = & \theta_2^T(t) \omega_2(t) + \theta_{20}(t) y(t) + \theta_3(t) r(t) \\ & + \theta_4^T(t) \omega_4(t) + \theta_5^T(t) \omega_5(t). \end{aligned} \quad (6.6)$$

$\theta_{20}(t) = 0$ in (6.6) if $a(s)$, $\Lambda(s)$ are defined by the controller II of Section V.

In view of (4.27), (5.4), (6.1)–(6.6), the tracking error equation now has the form

$$e(t) = \rho^* W_m(s) [\phi^T \omega](t) + d(t) \quad (6.7)$$

where $\rho^* = \theta_3^{*-1}$, $\phi(t) = \theta(t) + \theta^*$, $d(t) = \rho^* W_m(s) (1 - \theta_1^{*T}(a(s)/\Lambda(s))) [d_h](t)$, and

$$\theta(t) = (\theta_2^T(t), \theta_{20}(t), \theta_3(t), \theta_4^T(t), \theta_5^T(t), \theta_h^T(t))^T \quad (6.8)$$

$$\theta^* = (\theta_2^{*T}, \theta_{20}^*, \theta_3^*, \theta_4^{*T}, \theta_5^{*T}, \theta_h^{*T})^T \quad (6.9)$$

$$\omega(t) = (\omega_2^T(t), y(t), r(t), \omega_4^T(t), \omega_5^T(t), \omega_h^T(t))^T$$

The expression of the tracking error allows us to use adaptive linear control theory [11], [12] to develop an adaptive law for updating parameters of the adaptive hysteresis inverse and adaptive linear controller structure which make up our adaptive controller for the nonlinear plant (2.1).

We note that θ_{20}^* , θ_h^* , $\theta_{20}(t)$, $\theta_h(t)$, $\omega_h(t)$ are defined in different ways when the different designs of Sections IV and V are used; see (4.24)–(4.35), and (5.1).

Starting with tracking error equation (6.7), we can design adaptive schemes to update the parameter vector $\theta(t)$. Next, we present two such schemes.

Adaptive Scheme I: Letting $\rho(t)$ be the estimate of ρ^* and $L(s)$ be any Hurwitz polynomial of degree $n^* - 1$ such that $W_m(s)L(s)$ is strictly positive real, and introducing

$$\begin{aligned} \zeta(t) = & L^{-1}(s) [\omega](t), \\ \xi(t) = & \theta^T(t) \zeta(t) - L^{-1}(s) [\theta^T \omega](t) \end{aligned} \quad (6.11)$$

we define the estimation error $\epsilon(t)$ from

$$\epsilon(t) = e(t) + W_m(s)L(s) [\rho \xi - \alpha \epsilon (\zeta^T \zeta + \xi^2)](t) \quad (6.12)$$

where $\alpha > 0$, and update $\theta(t)$ and $\rho(t)$ from

$$\dot{\theta}(t) = -\text{sign}[k_p] \Gamma^T \zeta(t) \epsilon(t) - \Gamma \sigma(\theta, M_\theta, \sigma_0) \theta(t) \quad (6.13)$$

$$\dot{\rho}(t) = -\gamma \xi(t) \epsilon(t) - \gamma \sigma(\rho, M_\rho, \sigma_0) \rho(t) \quad (6.14)$$

where $\Gamma = \Gamma^T > 0$, $\gamma > 0$, and σ is the switching signal defined by (5.9).

We then have the following properties of this adaptive law.

Lemma 6.1. The adaptive law (6.13)–(6.14) guarantees

- 1) $\theta(t)$, $\rho(t)$, $\epsilon(t) \in L_\infty$;
- 2) for some constants $k_3 > 0$, $k_4 > 0$, and all $t_2 > t_1 \geq 0$,

$$\int_{t_1}^{t_2} \|\dot{\theta}(t)\|^2 dt \leq k_3 + \int_{t_1}^{t_2} \frac{k_4}{1 + \zeta^T(t) \zeta(t) + \xi^2(t)} dt \quad (6.15)$$

$$\begin{aligned} & \int_{t_1}^{t_2} \epsilon^2(t) (1 + \zeta^T(t) \zeta(t) + \xi^2(t)) dt \\ & \leq k_3 + \int_{t_1}^{t_2} \frac{k_4}{1 + \zeta^T(t) \zeta(t) + \xi^2(t)} dt \end{aligned} \quad (6.16)$$

Proof. We rewrite (6.7) as

$$\begin{aligned} e(t) = & W_m(s)L(s) [\rho^* (L^{-1}(s) [\theta^T \omega] - \theta^{*T} L^{-1}(s) [\omega] \\ & + \rho^* (1 - \theta_1^{*T} \frac{a(s)}{\Lambda(s)}) L^{-1}(s) [d_h])](t). \end{aligned} \quad (6.17)$$

Substituting (6.12) in (6.17), we have

$$\epsilon(t) = W_m(s)L(s) [\rho^* \phi^T \zeta + \psi \xi - \alpha \epsilon (\zeta^T \zeta + \xi^2) + \tilde{d}](t) \quad (6.18)$$

where $\tilde{d}(t) = \rho^* (1 - \theta_1^{*T}(a(s)/\Lambda(s))) L^{-1}(s) [d_h](t)$, $\psi(t) = \rho(t) - \rho^*$.

Let $W_m(s)L(s)$ have a controllable realization (A_m, B_m, C_m) and $e_r(t)$ be its state variable. With this notation, from (6.18), we obtain

$$\dot{e}_r(t) = A_m e_r(t) + B_m \nu(t), \quad \epsilon(t) = C_m e_r(t) \quad (6.19)$$

where

$$\nu(t) = \rho^* \phi^T(t) \zeta(t) + \psi(t) \xi(t) - \alpha(t)(\zeta^T(t) \zeta(t) + \xi^2(t)) + \bar{d}(t). \quad (6.20)$$

Strict positive realness of $W_m(s)J(s)$ implies that there exist constant matrices $Q = Q^T > 0$, $P = P^T > 0$, vector q , and scalar $\delta > 0$ such that $A_m^T P + P A_m = -q q^T + \delta Q$, $P B_m = C_m^T$. Hence, the time derivative of $V(\rho, \phi, \psi) = e_c^T P e_c + |\rho^*| \phi^T \Gamma^{-1} \phi + \gamma^{-1} \psi^2$ is

$$\begin{aligned} \dot{V} \leq & -\alpha_1 \epsilon^2(t)(1 + \zeta^T(t) \zeta(t) + \xi^2(t)) \\ & + \alpha_2 \frac{d^2(t)}{1 + \zeta^T(t) \zeta(t) + \xi^2(t)} \\ & - 2|\rho^*| \sigma_\theta(t) \phi^T(t) \theta(t) - 2\sigma_\rho(t) \psi(t) \rho(t) \end{aligned} \quad (6.21)$$

for some constants $\alpha_1 > 0$, $\alpha_2 > 0$. This, together with a similar argument to (5.14) and (5.15), proves (1) and (2). \square

Adaptive Scheme II: Instead of a choice of $L(s)$ of degree $n^* - 1$, we choose $L(s) = P_m(s) = W_m^{-1}(s)$, of degree n^* , to obtain the estimation error as

$$\epsilon(t) = \frac{e(t) + \rho(t) \xi(t)}{1 + \alpha(\zeta^T(t) \zeta(t) + \xi^2(t))} \quad (6.22)$$

where $\zeta(t)$ and $\xi(t)$ are defined in (6.11) with this new $L(s)$. We then use (6.13) and (6.14) with the new $\epsilon(t)$ defined in (6.22) to update the estimates $\theta(t)$ and $\rho(t)$.

For this scheme, we have the following result.

Lemma 6.2: The adaptive (6.13), (6.14), (6.23) guarantees

- 1) $\theta(t)$, $\dot{\theta}(t)$, $\rho(t)$, $\dot{\rho}(t)$, $\epsilon^2(t)(1 + \alpha(\zeta^T(t) \zeta(t) + \xi^2(t))) \in L_\infty$;
- 2) for some constant $k_5 > 0$, $k_6 > 0$, and all $t_2 \geq t_1 \geq 0$,

$$\int_{t_1}^{t_2} \|\dot{\theta}(t)\|^2 dt \leq k_5 + \int_{t_1}^{t_2} \frac{k_6}{1 + \alpha(\zeta^T(t) \zeta(t) + \xi^2(t))} dt \quad (6.23)$$

$$\begin{aligned} & \int_{t_1}^{t_2} \epsilon^2(t)(1 + \alpha(\zeta^T(t) \zeta(t) + \xi^2(t))) dt \\ & \leq k_5 + \int_{t_1}^{t_2} \frac{k_6}{1 + \alpha(\zeta^T(t) \zeta(t) + \xi^2(t))} dt. \end{aligned} \quad (6.24)$$

Proof: With (6.7), (6.11), and (6.22), we write the time derivative of the positive-definite function $V(\phi, \psi) = \frac{1}{2}(|\rho^*| \phi^T \Gamma^{-1} \phi + \gamma^{-1} \psi^2)$ where $\psi(t) = \rho(t) - \rho^*$, along (6.13) and (6.14), as

$$\begin{aligned} \dot{V}(t) \leq & -\frac{1}{2} \epsilon^2(t)(1 + \alpha(\zeta^T(t) \zeta(t) + \xi^2(t))) \\ & + \frac{d^2(t)}{2(1 + \alpha(\zeta^T(t) \zeta(t) + \xi^2(t)))} \\ & - |\rho^*| \sigma_\theta(t) \phi^T(t) \theta(t) \\ & - \sigma(\rho, M_\rho, \sigma_0) \psi(t) \rho(t). \end{aligned} \quad (6.25)$$

From this inequality, (6.13) and (6.14) with σ defined in (5.9), (1), and (2) follow. \square

Similar to Lemma 5.1, Lemmas 6.1 and 6.2 only show the boundedness of $\theta(t)$ and the L^2 properties (6.23)–(6.24) of the adaptive law (6.13)–(6.14). The signal boundedness of the adaptive control system with a general hysteresis is

still a research topic. However, for a wide class of hysteresis characteristics, we have the following stability results.

Theorem 6.1: If the hysteresis $H(\cdot)$ in (2.1) has two equal slopes, $m_t = m_b$, and (4.30), (4.31), (6.9), (6.10) are used for adaptation and implementation of the adaptive hysteresis inverse and the adaptive controller, then all closed-loop signals are bounded.

The proof for Theorem 6.1 is given in Appendix B.

This boundedness result applies to the symmetric hysteresis case where (4.34), (4.35), (6.9), and (6.10) are used for adaptation and implementation of the adaptive hysteresis inverse and the adaptive controller (6.6), with a reduced system order.

While analytically characterizing the tracking performance of the closed-loop system with an adaptive hysteresis inverse is an important future research topic, we will examine it by simulations and compare it with that of a control system without hysteresis inverse.

VII. SIMULATION RESULTS

In this section, we present an example with simulation results to illustrate the structure and effectiveness of the developed adaptive hysteresis control schemes.

We consider the plant with an unstable linear part $G(s) = 2.5/(s^2 + 2s - 4)$ and a symmetric hysteresis characteristic $H(m_t, c_t, m_b, c_b, m_r, c_r, m_l, c_l; \cdot)$ where $m_t = m_b = 1.8$, $m_r = m_l = 3$, $c_t = -c_b = 1.9$, $c_r = -c_l = 2.5$, and choose a model transfer function $W_m(s) = 1/(s^2 + 3s + 2)$.

We study the system performance in one of the five cases: a) a fixed linear controller without hysteresis inverse, for $G(s)$ known; b) an adaptive linear controller without hysteresis inverse, for $G(s)$ unknown; c) a fixed linear controller with a fixed inaccurate hysteresis inverse, for $G(s)$ known; d) a fixed linear controller with an adaptive hysteresis inverse, for $G(s)$ known; and e) an adaptive linear controller with an adaptive hysteresis inverse, for $G(s)$ unknown.

The controller a) has the structure

$$\begin{aligned} v(t) = & \theta_1^* \frac{1}{s+3} [v](t) + \theta_2^* \frac{1}{s+3} [y](t) \\ & + \theta_{20}^* y(t) + \theta_{31}^* r(t) \end{aligned} \quad (7.1)$$

where $\theta_1^* = -1$, $\theta_2^* = -0.4$, $\theta_{20}^* = -2.8$, $\theta_{31}^* = 0.4$, which are calculated from (5.3) with $G(s)$ known.

The controller b) is the adaptive version of a) for $G(s)$ unknown:

$$\begin{aligned} v(t) = & \theta_1(t) \frac{1}{s+3} [v](t) + \theta_2(t) \frac{1}{s+3} [y](t) \\ & + \theta_{20}(t) y(t) + \theta_3(t) r(t) \end{aligned} \quad (7.2)$$

where $\theta_1(t)$, $\theta_2(t)$, $\theta_{20}(t)$, $\theta_3(t)$ are updated from an adaptive law similar to (6.11)–(6.14) with a different $\theta(t) = (\theta_1(t), \theta_2(t), \theta_{20}(t), \theta_3(t))^T$, and with $\theta(0) = (-1.5, -0.8, -2, 0.8)^T$, $\rho(0) = 1.25$, and the choice of $L(s) = s + 2$, $\alpha = 1$, $\sigma_0 = 0.15$, $\Gamma = 10I$, $\gamma = 10$, $M_\theta = 5.53$, $M_\rho = 3$.

These two controllers, well known in the literature, ignore the existence of the hysteresis characteristic in the plant and have no compensation for the hysteresis $H(\cdot)$.

The controller c) is

$$v(t) = \bar{H}I(u_d(t)),$$

$$\bar{H}I(\cdot) = HI(\bar{m}_t, \bar{c}_t, \bar{m}_b, \bar{c}_b, \bar{m}_r, \bar{c}_r, \bar{m}_l, \bar{c}_l; \cdot) \quad (7.3)$$

$$u_d(t) = \theta_1^* \frac{1}{s+3} [u_d](t) + \theta_2^* \frac{1}{s+3} [y](t) + \theta_{20}^* y(t) + \theta_3^* r(t) \quad (7.4)$$

where $\bar{H}I(\cdot)$ is an accurate estimate of $HI(\cdot)$ defined in (3.2)–(3.6), with $\bar{m}_t = \bar{m}_b = 2.3$, $\bar{c}_t = -\bar{c}_b = 0.9$, $\bar{m}_r = \bar{m}_l = 2.5$, $\bar{c}_r = -\bar{c}_l = 0.5$, and θ_1^* , θ_2^* , θ_{20}^* , θ_3^* are the same as that in the controller a). This controller, with $\bar{H}I(\cdot)$, has a fixed but inaccurate compensation for the hysteresis $H(\cdot)$. The control law (7.4) was developed in (5.1).

The controller d) is with an adaptive hysteresis inverse for $G(s)$ known

$$v(t) = \widehat{H}I(u_d(t)), \quad \widehat{H}I(\cdot) = HI(\widehat{m}_t(t), \widehat{c}_t(t), \widehat{m}_b(t), \widehat{c}_b(t), \widehat{m}_r(t), \widehat{c}_r(t), \widehat{m}_l(t), \widehat{c}_l(t); \cdot) \quad (7.5)$$

where $u_d(t)$ is generated from the same structure as that in the controller c) [see (7.4)], and $\widehat{H}I(\cdot)$ is an adaptive estimate of $HI(\cdot)$, with initial parameter estimates $\widehat{m}_t(0) = \widehat{m}_b(0) = 2.3$, $\widehat{c}_t(0) = -\widehat{c}_b(0) = 0.9$, $\widehat{m}_r(0) = \widehat{m}_l(0) = 2.5$, $\widehat{c}_r(0) = -\widehat{c}_l(0) = 0.5$. The adaptive law for $\theta_h(t) = (\widehat{m}_t(t), \widehat{c}_t(t), \widehat{m}_r(t), \widehat{m}_l(t), \widehat{c}_r(t))^T$ is (5.8) with $\sigma_0 = 0.15$, $\Gamma_h = 10I$, $M_h = 11$.

The controller (e) is with an adaptive hysteresis inverse for $G(s)$ unknown

$$v(t) = \widehat{H}I(u_d(t)), \quad \widehat{H}I(\cdot) = HI(\widehat{m}_t(t), \widehat{c}_t(t), \widehat{m}_b(t), \widehat{c}_b(t), \widehat{m}_r(t), \widehat{c}_r(t), \widehat{m}_l(t), \widehat{c}_l(t); \cdot) \quad (7.6)$$

$$u_d(t) = \theta_2(t) \frac{1}{s+3} [y](t) + \theta_{20}(t) y(t) + \theta_3(t) r(t) + (\theta_{41}(t), \theta_{42}(t), \theta_{43}(t), \theta_{44}(t)) \frac{1}{s+3} [\omega_h](t) + \theta_5(t) \omega_5(t) \quad (7.7)$$

where $\widehat{H}I(\cdot)$ is an adaptive hysteresis inverse with the same initial parameter estimates as that in the controller d), and the adaptive law for $\theta(t) = (\theta_2(t), \theta_{20}(t), \theta_3(t), \theta_4^T(t), \theta_5(t), \theta_h^T(t))^T$ is (6.11)–(6.14) with the choice of $L(s) = s+2$, $\alpha = 1$, $\sigma_0 = 0.15$, $\Gamma = 10I$, $\gamma = 10$, $M_\theta = 17.38$, $M_\rho = 3$, $\rho(0) = 1.25$, and $\theta_2(0) = -0.8$, $\theta_{20}(0) = -2$, $\theta_3(0) = 0.8$, $\theta_4(0) = (3.45, 3.75, 1.35, 0.75)^T$, $\theta_5(0) = -1.5$. The control law (7.7) was developed in (6.6).

Our simulations indicate

1) if either a fixed or adaptive linear controller without a hysteresis inverse is applied to this plant, then the tracking error remains significant for large t ;

2) an inaccurate fixed hysteresis inverse can reduce the tracking error;

3) significant improvements of system tracking performance are achieved with the use of an adaptive hysteresis inverse either with a fixed linear controller, for $G(s)$ known, or with an adaptive linear controller, for $G(s)$ unknown.

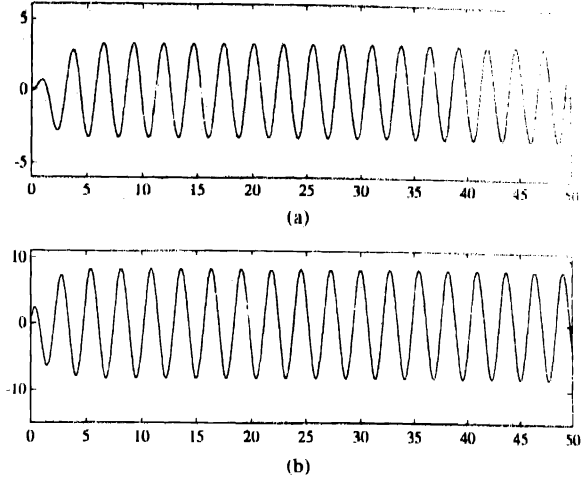


Fig. 4. System responses: fixed controller without hysteresis inverse for $G(s)$ known.

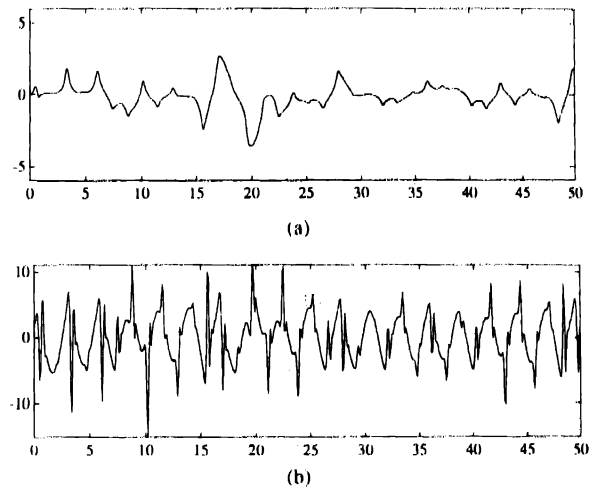


Fig. 5. System responses: adaptive controller without hysteresis inverse for $G(s)$ unknown.

Some typical system responses for $r(t) = 12.7 \sin(2.3t)$ are shown in Figs. 4–8 for the above five controllers, respectively. It is clear that with an adaptive hysteresis inverse, the tracking error is significantly reduced with less control effort (the control signal is smoother and smaller).

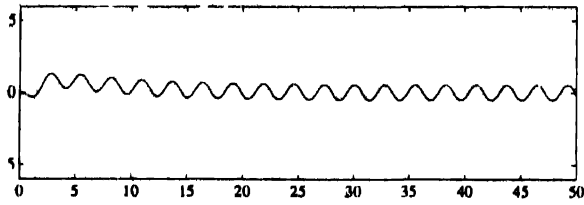
VIII. CONCLUSIONS

Our construction of an adaptive hysteresis inverse, reparameterization of an adaptive linear controller, and choice of suitable error models have led to several adaptive control schemes applicable to plants with unknown hystereses. Simulation results indicate that these adaptive inverse schemes promise to significantly improve system performance.

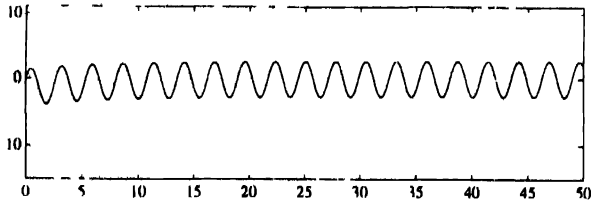
APPENDIX A

PROOF OF THEOREM 5.1

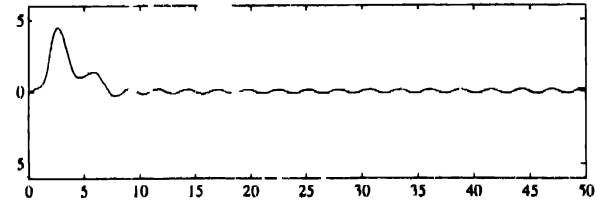
From Lemma 5.1 and the condition 7), we have that $\theta_h(t) \in L_\infty$ and $\widehat{m}_t(t) \geq m_{t1} > 0$. Introducing $m = m_t = m_b$, $\widehat{m}_t(t) = \widehat{m}_b(t)$, and using $u(t) = H(v(t))$, $v(t) =$



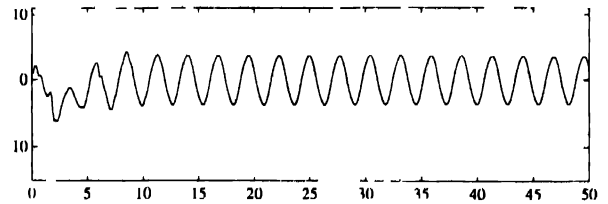
(a)



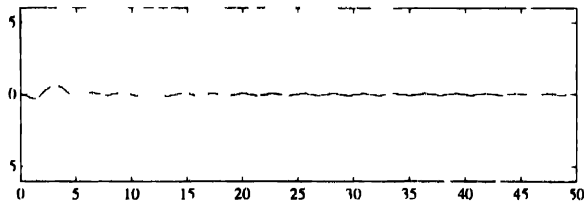
(b)

Fig 6 System responses fixed controller with inaccurate hysteresis inverse for $G(s)$ known

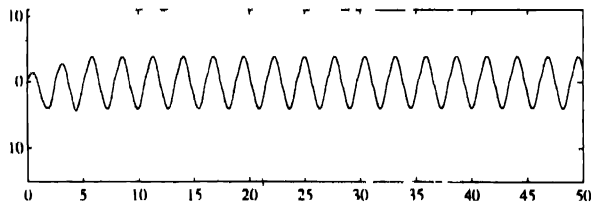
(a)



(b)

Fig 8 System responses adaptive controller with adaptive hysteresis inverse for $G(s)$ unknown

(a)



(b)

Fig 7 System responses fixed controller with adaptive hysteresis inverse for $G(s)$ known

$\hat{H}(u_d(t))$ we express

$$u(t) = \frac{m}{m(t)} u_d(t) + d_2(t) \quad (\text{A } 1)$$

$$u(t) = mv(t) + \bar{d}_2(t) \quad (\text{A } 2)$$

for some bounded $d_2(t)$, $\bar{d}_2(t)$. We then define fictitious signals $z_0(t)$, $z_1(t)$ filters $K_1(s)$, $K(s)$ as

$$z_0(t) = \frac{1}{s + a_0} [u](t), \quad \bar{z}_1(t) = \frac{1}{s + a_0} [y](t), \quad a_0 > 0 \quad (\text{A } 3)$$

$$K_1(s) = 1 - K(s), \quad K(s) = \frac{a^n}{(s + a)^n} \quad a > 0 \quad (\text{A } 4)$$

and use (A.3), (A.4), and (2.1) to obtain

$$z_0(t) + a_0 K_1(s) [z_0](t) - K_1(s) [u](t) = K(s) G^{-1}(s) [z](t) \quad (\text{A } 5)$$

Using (A.1) and (A.3) in (5.1), we have

$$\begin{aligned} u(t) = & \frac{m}{m(t)} \theta_1^* I \frac{a(s)}{\Lambda(s)} \frac{m(\cdot)}{m} (s + a_0) [z_0](t) + \\ & \frac{m}{m(t)} \theta_2^* I \frac{a(s)}{\Lambda(s)} (s + a_0) [z_1](t) \\ & + \frac{m}{m(t)} \theta_{20}^* (s + a_0) [z](t) + \frac{m}{m(t)} \theta_1^* I(t) \\ & + \left(1 - \frac{m}{m(t)} \theta_1^* I \frac{a(s)}{\Lambda(s)} \frac{m(\cdot)}{m} \right) [d_2](t) \end{aligned} \quad (\text{A } 6)$$

Substituting this expression for $u(t)$ in (A.5), we see that $\bar{z}_1(t)$ and $z_0(t)$ are related by

$$\begin{aligned} & \left(1 + K_1(s) \left(a_0 - \frac{m}{m(\cdot)} \theta_1^* I \frac{a(s)}{\Lambda(s)} \frac{m(\cdot)}{m} (s + a_0) \right) \right) [z_0](t) \\ & = \left(K(s) G^{-1}(s) + K_1(s) \frac{m}{m(\cdot)} \theta_2^* I \frac{a(s)}{\Lambda(s)} (s + a_0) \right. \\ & \quad + K_1(s) \frac{m}{m(\cdot)} \theta_{20}^* (s + a_0) [z](t) \\ & \quad + K_1(s) \left[\frac{m}{m(\cdot)} \theta_3^* I \right] (t) \\ & \quad \left. + K_1(s) \left(1 - \frac{m}{m(\cdot)} \theta_1^* I \frac{a(s)}{\Lambda(s)} \frac{m(\cdot)}{m} \right) [d_2](t) \right) \end{aligned} \quad (\text{A } 7)$$

Consider a linear operator $T(s, t)$ with input $r(t)$. We define $T(s, t)$ as a stable and proper operator if $\|T(s, \cdot)[r](t)\| \leq \beta_1 \int_0^t e^{-\alpha(t-\tau)} \|r(\tau)\| d\tau + \beta_2 \|r(t)\|$ for some constants $\beta_1 \geq 0$, $\beta_2 \geq 0$, and $\alpha > 0$ all $t \geq 0$ and any $r(t)$. The operator $T(s, t)$ is stable and strictly proper if $\|T(s, \cdot)[r](t)\| \leq \beta_1 \int_0^t e^{-\alpha(t-\tau)} \|r(\tau)\| d\tau$ for some constants $\beta_1 \geq 0$ and $\alpha > 0$, all $t \geq 0$ and any $r(t)$. With this definition, the facts that $\hat{m}(t)$, $m(t) \in L_\infty$, $K_1(s)$ is strictly stable, and

$$\begin{aligned} & \frac{m}{m(t)} \theta_1^* I \frac{a(s)}{\Lambda(s)} \frac{m(\cdot)}{m} (s + a_0) [z_0](t) \\ & = \frac{m}{m(t)} \theta_1^* I \frac{a(s)}{\Lambda(s)} \left[s \left[\frac{m}{m} z_0 \right] - s \left[\frac{\hat{m}}{m} \right] z_0 + a_0 \frac{m}{m} z_0 \right] (t) \end{aligned} \quad (\text{A } 8)$$

$$K_1(s) \frac{m}{m(\cdot)} \theta_{20}^*(s+a_0)[z](t) \\ = K_1(s) \left[s \left[\frac{m}{m(\cdot)} \theta_{20}^* z \right] - s \left[\frac{m}{m(\cdot)} \theta_{20}^* \right] z + a_0 \frac{m}{m(\cdot)} \theta_{20}^* z \right] (t) \quad (\text{A } 9)$$

imply that $(m/m(\cdot))\theta_1^{*T}(a(s)/\Lambda(s))(m(t)/m)(s+a_0)$ and $K_1(s)(m/m(\cdot))\theta_{20}^*(s+a_0)$ are stable and proper

By definition, the impulse response function $k_1(t)$ of $K_1(s)$ satisfies $\int_0^\infty |k_1(t)|dt = n^*/a$. Hence there exists $a^0 > 0$ such that for any finite $a > a^0$ the operator

$$T_0(s, t) \triangleq \left(1 + K_1(s) \left(a_0 - \frac{m}{m(\cdot)} \theta_1^{*T} \frac{a(s)}{\Lambda(s)} \frac{m(t)}{m} (s+a_0) \right) \right)^{-1}$$

is stable and proper. For a fixed $a > a^0$ (A 7) implies that

$$z_0(t) = I_1(s, t)[z](t) + d_1(t) \quad (\text{A } 10)$$

where $I_1(s, t)$ is a stable and proper operator, and $d_1(t)$ is a bounded signal due to $r(t)$, $d_2(t)$

As in the swapping lemma [12] we let (A_f, B_f, C_f) be a minimal realization of $I(s)$ define $W(s) = (C_f(sI - A_f))^{-1}$, $W_t(s) = (sI - A_f)^{-1}B_f$ and express $\xi_h(t)$ as

$$\xi_h(t) = W(s, t) \left[W_t(s)(s+a_0) \frac{1}{s+a_0} [\omega_h^T(t)] \theta_h \right] (t) \quad (\text{A } 11)$$

Using (4 13) and (4 31) we rewrite $\omega_h(t)$ as

$$\omega_h(t) = (-\dot{\chi}_t(t) + (\hat{\chi}_t(t) + \chi_t(t))\dot{\chi}_t(t) - \chi_t(t) - \hat{\chi}_t(t)) \\ \hat{\chi}_t(t)\dot{\chi}_t(t) - \hat{\chi}_t(t) - \chi_t(t)\dot{\chi}_t(t) - \hat{\chi}_t(t))^T \quad (\text{A } 12)$$

whose components except for $-\dot{\chi}_t(t)$ are all bounded. Using (A 2) and (A 3) we note

$$\frac{1}{s+a_0} [\omega_h](t) = \frac{1}{m} (v_0(t), 0, 0)^T + d_3(t) \quad (\text{A } 13)$$

where $d_3(t)$ is a bounded vector signal

Filtering both sides of (5 7) by $1/(s+a_0)$ and using (A 3) again, we obtain

$$z(t) = \frac{1}{s+a_0} [y_m](t) + \frac{1}{s+a_0} [e_h - \xi_h](t) \quad (\text{A } 14)$$

The inequality $|e_h(t)| < (|e_h(t)|/\sqrt{1+\zeta_h^T(t)\zeta_h(t)+\xi_h^2(t)}) (1+\|\zeta_h(t)\|+|\xi_h(t)|)$ and (A 9)–(A 14) imply

$$|z(t)| \leq v_0(t) + I_2(s, t)[v_1 I_3(s, t)[z]](t) \quad (\text{A } 15)$$

where $v_0(t) \in L_\infty$ and $v_1(t) \in L_\infty$ is such that

$$\int_{t_1}^t v_1^2(t) dt \leq \bar{k}_1 + \int_{t_1}^t \frac{\bar{k}_2}{1+\zeta_h^T(t)\zeta_h(t)+\xi_h^2(t)} dt \quad (\text{A } 16)$$

for some constants $\bar{k}_1 > 0$, $\bar{k}_2 > 0$ and any $t_2 \geq t_1 \geq 0$. The operator $T_2(s, t)$ is stable and strictly proper, while the operator $T_3(s, t)$ is stable and proper and has a nonnegative impulse response

If $\zeta_h(t)$, $\xi_h(t)$ are bounded, then from Lemma 5.1 and (5 7), we see that $e_h(t)$, $y(t)$ are bounded, from (A 3), (A 10) that $z(t)$, $z_0(t)$ are bounded, and from (A 6), (A 8), and (A 1) that $u(t)$ and $u_d(t)$ are bounded. Thus all closed-loop signals

are bounded. On the other hand, if $\zeta_h^T(t)\zeta_h(t)$ is unbounded, then the smallness of $v_1(t)$ in the second term results in the boundedness of $z(t)$ given by (A 15). This in turn implies that $z_0(t)$ in (A 10) is bounded so that (A 6), $u_d(t)$ in (A 1), $\omega_h(t)$ in (4 13), $\xi_h(t)$ in (5 6), in (5 7) are bounded. Hence we conclude that all closed-loop signals are bounded.

APPENDIX B PROOF OF THEOREM 6.1

A Proof for Adaptive Scheme I

Using (A 1), (A 3), (A 11), (A 12), and (6 3) we express

$$\theta_4^T(t)\omega_4(t) = \theta_a^T(t) \frac{a(s)}{\Lambda(s)} (s+a_0)[z_0](t) + d_5(t) \quad (\text{B } 1)$$

for some bounded $\theta_a(t) \in R^{n-1}$ for the controller I ($\theta_a(t) \in R^n$ for the controller II) of Section V and some bounded $d_5(t)$. Using $z_0(t)$, $K_1(s)$, $K(s)$ defined in (A 3)–(A 4) we obtain

$$z_0(t) + a_0 K_1(s)[z_0](t) - K_1(s)[u](t) \\ = K(s)G^{-1}(s) \frac{1}{s+a_0} [y](t) \quad (\text{B } 2)$$

Using (A 1)–(6 6)–(B 1), we obtain

$$u(t) = \frac{m}{m(t)} \theta_a^T(t) \frac{a(s)}{\Lambda(s)} (s+a_0)[z_0](t) \\ + \frac{m}{m(t)} \theta_2^T(t) \frac{a(s)}{\Lambda(s)} [y](t) + \frac{m}{m(t)} \theta_{20}^T(t) y(t) \\ + \frac{m}{m(t)} \theta_3(t) r(t) + \frac{m}{m(t)} d_5(t) + d_2(t) \quad (\text{B } 3)$$

From (B 2) and (B 3) it follows that

$$\left(1 + K_1(s) \left(a_0 - \frac{m}{m(\cdot)} \theta_a^T(s) \frac{a(s)}{\Lambda(s)} (s+a_0) \right) \right) [z_0](t) \\ = \left(K(s)G^{-1}(s) \frac{1}{s+a_0} + K_1(s) \frac{m}{m(\cdot)} \theta_2^T(s) \frac{a(s)}{\Lambda(s)} \right. \\ \left. + K_1(s) \frac{m}{m(\cdot)} \theta_{20}^T(s) \right) [y](t) + K_1(s) \left[\frac{m}{m} \theta_3 r \right] (t) \\ + K_1(s)[d_2](t) + K_1(s) \left[\frac{m}{m} d_5 \right] (t) \quad (\text{B } 4)$$

Similar to (A 9), we see from (B 4) with a sufficiently large $a > 0$ in $K(s)$, $K_1(s)$ that

$$z_0(t) = T_1(s, t)[y](t) + \bar{d}_3(t) \quad (\text{B } 5)$$

for some $T_1(s, t)$ stable and proper and some $\bar{d}_3(t)$ bounded

Similar to (A 10) with (A_f, B_f, C_f) being a minimal realization of $1/L(s)$ and $W(s) = C_f(sI - A_f)^{-1}$, $W_t(s) = (sI - A_f)^{-1}B_f$ we express $\xi(t)$ as

$$\xi(t) = W(s, t)[W_t(s)[\omega^T(t)]\theta](t) \quad (\text{B } 6)$$

Using (A 11), (6 3), (6 4), (5 2), (B 5), in (6 10), we see that

$$\omega(t) = \bar{T}_1(s, t)[y](t) + \bar{d}_4(t) \quad (\text{B } 7)$$

for some $\bar{T}_1(s, t)$ stable and proper, and some $\bar{d}_4(t)$ bounded

By rewriting (6.12), we get

$$y(t) = y_m(t) + e(t) - W_m(s)L(s) [\rho\xi - \alpha\xi(1 + \zeta^T\zeta + \xi^2) + \alpha e](t) \quad (B.8)$$

Similar to (A.15), (6.11), (B.6)–(B.8), and Lemma 6.1 imply

$$|y(t)| \leq r_0(t) + I_2(s)[r_1 T_3(s)[|y|]](t) \quad (B.9)$$

where $r_0(t) \in L_\infty$, and $r_1(t)$ is such that

$$\int_{t_1}^{t_2} r_1^2(t) dt \leq \bar{k}_3 + \int_{t_1}^{t_2} \frac{\bar{k}_4}{1 + \zeta^T(t)\zeta(t) + \xi^2(t)} dt \quad (B.10)$$

for some constants $\bar{k}_3 > 0$, $\bar{k}_4 > 0$ and any $t_2 \geq t_1 \geq 0$. The operator $T_2(s, t)$ is stable and strictly proper while the operator $T_3(s, t)$ is stable and proper and has a nonnegative impulse response. From here on, the closed-loop signal boundedness follows from a contradiction argument which is similar to that in the proof of Theorem 5.1.

B Proof for Adaptive Scheme II

Using (A.1), (6.6), (B.1), we obtain

$$\begin{aligned} u(t) = & \frac{m}{m(t)} \theta_a^T(t) \frac{a(s)}{\Lambda(s)} (s + a_0)[z_0](t) \\ & + \frac{m}{m(t)} \theta_2^T(t) \frac{a(s)}{\Lambda(s)} (s + a_0)[z](t) \\ & + \frac{m}{m(t)} \theta_{20}(t) (s + a_0)[\sim](t) \\ & + \frac{m}{m(t)} \theta_3(t) \tau(t) + \frac{m}{m(t)} d_5(t) + d_2(t) \end{aligned} \quad (B.11)$$

From (A.5) and (B.11) it follows that

$$\begin{aligned} & \left(1 + K_1(s) \left(a_0 - \frac{m}{m(t)} \theta_a^T(t) \frac{a(s)}{\Lambda(s)} (s + a_0) \right) \right) [z_0](t) \\ = & \left(K(s) G^{-1}(s) + K_1(s) \frac{m}{m(t)} \theta_2^T(t) \frac{a(s)}{\Lambda(s)} (s + a_0) \right. \\ & \left. + K_1(s) \frac{m}{m(t)} \theta_{20}(t) (s + a_0) \right) [z](t) \\ & + K_1(s) \left[\frac{m}{m(t)} \theta_3(t) \right] (t) + K_1(s) [d_2](t) \\ & + K_1(s) \left[\frac{m}{m(t)} d_5 \right] (t) \end{aligned} \quad (B.12)$$

where $K_1(s) \frac{m}{m(t)} \theta_{20}(t) (s + a_0)$ is stable and proper because

$$\begin{aligned} & K_1(s) \frac{m}{m(t)} \theta_{20}(t) (s + a_0) [z](t) \\ = & K_1(s) \left[s \left[\frac{m}{m(t)} \theta_{20}(t) \right] - s \left[\frac{m}{m(t)} \theta_{20}(t) \right] z + a_0 \frac{m}{m(t)} \theta_{20}(t) \right] (t) \end{aligned} \quad (B.13)$$

and $m(t)$, $\theta_a(t)$, $\theta_{20}(t)$, $\theta_3(t)$ are all bounded.

The remaining part of the proof is analogous to (A.9)–(A.14), followed by a contradiction argument similar to that used in the proof of Theorem 5.1.

ACKNOWLEDGMENT

We are thankful to J. Winkelman and D. Rhode of Ford Motor Company for stimulating this research, and to D. Recker of the University of Illinois for many helpful discussions.

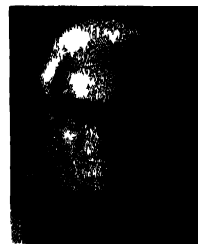
REFERENCES

- [1] I. O. Chua and S. C. Bass, "A generalized hysteresis model," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 36–48, Jan. 1972.
- [2] G. J. Thaler and M. P. Paskel, *Analysis and Design of Nonlinear Feedback Control Systems*. New York: McGraw-Hill, 1962.
- [3] M. A. Krasnosel'skii and A. V. Pokrovskii, *Systems with Hysteresis*. Berlin: Springer-Verlag, 1983.
- [4] N. D. Vaughan and J. B. Gamble, "The modelling and simulation of a proportional solenoid valve," presented at the Winter Ann. Meet. Amer. Soc. Mech. Eng., Dallas, TX, Nov. 1990.
- [5] I. D. Mayergoyz, *Mathematical Models of Hysteresis*. Berlin: Springer-Verlag, 1991.
- [6] A. Visintin, "Mathematical models of hysteresis," in *Topics in Non-smooth Mechanics*, J. J. Moreau, P. D. Panagiotopoulos, and G. Strang, Eds., Berlin: Birkhäuser Verlag, 1988, pp. 295–326.
- [7] D. A. Recker, P. V. Kokotovic, D. S. Rhode, and J. R. Winkelman, "Adaptive nonlinear control of systems containing a dead zone," in *Proc. 30th IEEE Conf. Decis. Contr.*, Brighton, England, 1991, pp. 2111–2115.
- [8] G. Tao and P. V. Kokotovic, "Adaptive control of plants with unknown dead zones," in *Proc. 1992 Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 2710–2714.
- [9] ———, "Adaptive control of systems with backlash," *Automatica*, vol. 29, pp. 323–335, Mar. 1993.
- [10] P. A. Ioannou and K. S. Tsakalis, "A robust direct adaptive controller," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 1033–1043, Nov. 1986.
- [11] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [12] S. S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence, and Robustness*. Englewood Cliffs, NJ: Prentice-Hall, 1989.



Gang Tao (S'84–M'89) received the B.S. degree from the University of Science and Technology of China in 1982, and the Ph.D. degree from the University of Southern California in 1989, both in electrical engineering.

He was a Visiting Assistant Professor at the Washington State University from 1989 to 1991, and an Assistant Research Engineer and a Lecturer at the University of California at Santa Barbara from 1991 to 1992. Since September 1992, he has been an Assistant Professor with the Department of Electrical Engineering, University of Virginia. His main research area is adaptive control, and currently he is working on the adaptive control of systems with nonsmooth nonlinearities, and on developing new adaptive algorithms for control and estimation.



Petar V. Kokotovic (SM 74–F 80) has been active for more than 30 years as a control engineer, researcher, and educator, first in his native Yugoslavia, and then from 1966 through 1990 at the University of Illinois, where he held the endowed Grainger Chair. Since 1991, he has been Co-Director (with A. J. Laub) of the newly formed Center for Control Engineering and Computation at the University of California, Santa Barbara. He has coauthored eight books and numerous articles contributing to sensitivity analysis, singular perturbation methods,

and robust adaptive and nonlinear control. He is also active in industrial applications of control theory. As a consultant to Ford, he was involved in the development of the first series of automotive computer controls, and at General Electric, he participated in large-scale systems studies.

Dr. Kokotovic received the 1990 Quazza Medal, the 1983 and 1993 Outstanding IEEE Transactions Paper Awards, and presented the 1991 Bode Prize Lecture. He is the recipient of the 1995 IEEE Control Systems Award.

Robust Stability Under a Class of Nonlinear Parametric Perturbations

Minyue Fu, *Member, IEEE*, Soura Dasgupta, *Senior Member, IEEE*, and Vincent Blondel

Abstract—This paper considers the robust stability verification of linear time-invariant systems admitting a class of nonlinear parametric perturbations. The general setting is one of determining the closed-loop stability of systems whose open-loop transfer functions consist of powers, products, and ratios of polytopes of polynomials. Apart from this general setting, two special cases of independent interest are also considered. The first special case concerns uncertainties in the open-loop gain and real poles and zeros, while the second special case treats uncertainties in the open-loop gain and complex poles and zeros. In light of the zero exclusion principle, robust stability is equivalent to zero exclusion of the value sets of the system characteristic function (a value set consists of the values of the characteristic functions at a fixed frequency). The main results of this paper are as follows. 1) The value set of the characteristic function at each fixed frequency is determined by the edges and some frequency-dependent internal line segments. 2) Consequently, Hurwitz invariance verification simplifies to that of checking certain continuous scalar functions for avoidance of the negative real axis. 3) For the case of real zero-pole-gain variations, the critical lines are all frequency independent, and therefore, the determination of the robust stability is even simpler. 4) For the case of complex zero-pole-gain variations, the critical internal lines are shown to be either frequency independent or to be confined in certain (two-dimensional) planes or (three-dimensional) boxes.

1 INTRODUCTION

THE following problem is of interest in the robust stability verification of linear time-invariant control systems depicted in Fig. 1. Suppose we are given a stability region \mathcal{D} and a family of open-loop transfer functions parameterized by a real vector γ

$$T(\gamma) = \{t(s, \gamma) \mid \gamma \in \Gamma\} \quad (1.1)$$

where $t(s, \gamma)$ is the transfer function of the plant and controller, and Γ is a connected set in \mathbb{R}^N . Determine as simply

Manuscript received October 12, 1989; revised September 14, 1992 and February 28, 1994. Recommended by Associate Editor, C. L. DeMarco. This work was supported in part by NSF Grants MIP-9001170, ECS-9211593, and ECS 9350346 and by a grant from the Université Catholique de Louvain-la-Neuve.

M. Fu was with the Laboratoire d'Automatique et d'Analyse des Systèmes, Université Catholique de Louvain, Louvain-la-Neuve, Belgium. He is now with the Department of Electrical and Computer Engineering, University of Newcastle, Newcastle, N.S.W. 2308, Australia.

S. Dasgupta was on leave at the Université Catholique de Louvain, Louvain-la-Neuve, Belgium. He is now with the Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA.

V. Blondel was with the Université Catholique de Louvain, Louvain-la-Neuve, Belgium. He is now with the Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, KTH, 100 44 Stockholm, Sweden.

IEEE Log Number 9407561

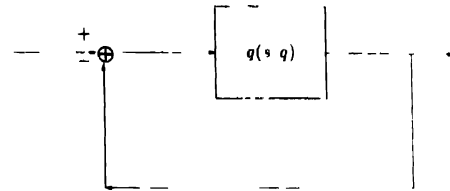


Fig. 1 Closed-loop uncertain system

as possible if all members of the family of the corresponding characteristic functions

$$H(\Gamma) = \{h(s, \gamma) = 1 + t(s, \gamma) \mid \gamma \in \Gamma\} \quad (1.2)$$

have all zeros contained entirely in \mathcal{D} (i.e., the family is \mathcal{D} -stable invariant). Generally, the transfer function coefficients depend nonlinearly on γ .

One approach to this problem is to treat it in its broadest generality, as is done in [1], [2] where a very broad class of $H(\Gamma)$ is considered. Alternatively, one can consider particular parameterizations reflecting specific forms of structural information supplied by the modeling process. This allows formulation of stability verification schemes which are computationally less demanding. Examples of this approach include [3], which considers a family of polynomials admitting independent variations in the coefficients, [4]–[6], which account for affinely dependent variations, and [7]–[9], which consider multilinear dependence (see [10], [11] for surveys). Each of [3]–[9], exploits the underlying structural information and demonstrates consequent simplifications.

This paper adopts the second approach by focusing on a special class of nonlinear parametric dependence. To keep the presentation simple, only Hurwitz invariance is investigated (i.e., \mathcal{D} is the open left-half plane), although the results do, in fact, generalize to more general stability regions. Specifically, the family of characteristic functions to be considered in this paper admits the following form

$$h(s, \gamma) = 1 + g_0(s) \frac{\prod_{i=1}^m (p_{i0}(s) + \gamma_i^T P_i(s))^\mu}{\prod_{i=m+1}^n (p_{i0}(s) + \gamma_i^T P_i(s))^\nu} \quad (1.3)$$

where $g_0(s)$ and the $p_{i0}(s)$ are real scalar rational functions and polynomials in s , respectively, $\gamma_i \in \Gamma_i \subset \mathbb{R}^N$ represent a partition of γ , i.e.,

$$\gamma = (\gamma_1^T, \gamma_2^T, \dots, \gamma_n^T)^T \quad (1.4)$$

and

$$\Gamma = \Gamma_1 \times \Gamma_2 \times \dots \times \Gamma_n \quad (1.5)$$

$P_i(s)$ are real vector polynomials with dimension N_i , and μ_i and ν_i are fixed positive exponents. The quantities $g_0(s)$, $p_{i0}(s)$, $P_i(s)$, n , N_i , μ_i , ν_i , and Γ are assumed known. The j th element of γ_i [respectively, $P_i(s)$] is denoted by γ_{ij} [respectively, $P_{ij}(s)$]. Since Γ is an axis parallel box, each γ_{ij} varies independently of the others within an interval

$$\gamma_{ij}^- \leq \gamma_{ij} \leq \gamma_{ij}^+. \quad (1.6)$$

Thus, each factor $(p_{i0}(s) + \gamma_i^T P_i(s))$ forms a polytope of polynomials as γ_i varies in Γ_i . Notice that a polytope can simplify to model an uncertain gain in the open-loop transfer function. Furthermore, in the case where the exponents μ_i and ν_i are restricted to be +1, the robust stability verification of (1.3) is equivalent to that of a subclass of the multilinear problem, i.e., the characteristic polynomial associated with (1.3) depends on γ_{ij} in a multilinear fashion. To simplify our notation, we rewrite (1.3) as follows:

$$h(s, \gamma) = 1 + g_0(s) \prod_{i=1}^n (p_{i0}(s) + \gamma_i^T P_i(s))^{k_i} \quad (1.7)$$

where k_i are allowed to be nonzero integers (either positive or negative).

There are several situations in which the setting of (1.7) becomes important. By way of background, we cite the results of [4], where the setting considered translates to one involving an uncertain plant having numerator and denominator polynomials lying in polytopic sets. Essentially, [4] asks the following question: Given a fixed controller $g_0(s)$, how can one determine if the uncertain closed loop is stable? Thus, the problem considered in [4] can be viewed as a subclass of (1.7), with $n = 2$, $k_1 = 1$, and $k_2 = -1$. In many applications involving process control, the overall plant is itself a cascade of several subplants, such that physical uncertain parameters of a given subplant enjoy physical independence from those of the other subplants. Now, if one models each uncertain subplant individually as lying in sets analogous to those in [4], one readily obtains a special case of the structure exhibited in (1.7).

To further illustrate the scope of (1.7), two more examples are considered. The first example is of a plant with independent real zero, pole and gain variations and is as follows

$$h(s, \gamma) = 1 + dg_0(s) \prod_{i=1}^{n-1} (s + \lambda_i)^{k_i} \quad (1.8)$$

where d and λ_i are uncertain real parameters lying in given bounds. In this case, $g_0(s)$ can be a given controller, d the gain, and λ_i the zeros and poles of the plant whenever the k_i are, respectively, positive or negative. The uncertainty assumes that the λ_i vary independently within given bounds. The objective is to verify if $g_0(s)$ stabilizes the plant for all possible parameter variations. Therefore, under the assumption of no zero-pole cancellation (which is trivial to check in this example), the closed-loop system associated with (1.8) is robustly stable if and only if the corresponding $H(\Gamma)$ is Hurwitz invariant. Note that the terms involving cases where $k_i \neq \pm 1$ reflect structural preservation of multiplicities.

To allow for complex zero and pole variations, one may include factors of the forms $(s^2 + a_i s + b_i)^{k_i}$, with a_i and b_i also varying independently in intervals. The version of (1.7) given in (1.9) below will be referred to as the complex zero-pole-gain variation problem:

$$h(s, \gamma) = 1 + dg_0(s) \prod_{i=1}^{\tau} (s + \lambda_i)^{k_i} \prod_{j=\tau+1}^{n-1} (s^2 + a_j s + b_j)^{k_j} \quad (1.9)$$

with

$$\gamma = (\lambda_1, \lambda_2, \dots, \lambda_{\tau}, a_{\tau+1}, b_{\tau+1}, \dots, a_{n-1}, b_{n-1}, d)^T. \quad (1.10)$$

Throughout this paper, we adopt the following assumptions on the function $h(s, \gamma)$ in (1.7).

Assumption 1.1: The function $h(s, \gamma)$ has no unstable zero-pole cancellation for any $\gamma \in \Gamma$.

Assumption 1.2: Continuous variations of γ result in continuous changes in the zeros of $h(s, \gamma)$.

Assumption 1.3: The function $h(s, \gamma)$ has no purely imaginary poles for any $\gamma \in \Gamma$.

We note that Assumption 1.1 is essential to assure the internal stability of the closed-loop system. However, violation of this assumption will not cause any difficulty for constructing the value set $H(j\omega, \Gamma)$. Further, in light of the recent results of [20], the verification of this assumption is a relatively simple matter.

Assumption 1.2, on the other hand, is a standard one that requires the leading coefficient of the overall numerator polynomial in (1.7) not to equal zero within the specified uncertainty bounds.

Finally, Assumption 1.3 can be simply tested through a series of straightforward algebraic [18] techniques.

A. Approach and Main Results

As in [8], [9], [12], [13], we follow the so-called value set approach to robust stability analysis. For the family of rational functions (1.2), the value set at a frequency ω is defined as

$$H(\omega, \Gamma) := \{h(j\omega, \gamma) : \gamma \in \Gamma\}, \quad (1.11)$$

and its boundary will be denoted by $\partial H(\omega, \Gamma)$.

Our approach exploits a slight variation of the zero exclusion principle (see, e.g., [9]). Under assumptions (1.1), (1.2), and the fact that the value set changes continuously with ω , this principle reduces to the following conditions as being necessary and sufficient for the Hurwitz invariance of $H(\Gamma)$:

- 1) at least one member of $H(\Gamma)$ is Hurwitz, and
- 2) $0 \notin \partial H(\omega, \Gamma)$, $\forall \omega \in \mathbb{R}$, where ∂ denotes "the boundary of."

There are several appealing qualities of the value set approach. First, it provides a unifying framework within which most of the currently known robust stability results can be understood. Second, it can be used to obtain simple and transparent proofs of these results (see, for example, [13], [14], [19] for proofs of [3], [4], [15]). Furthermore, while for certain special classes of uncertainties the robust stability of a family of polynomials reduces to that of some easily characterizable members of this family, the same does not hold for general

parameterizations. For example, when the uncertainty set is one of multiaffinely parameterized polynomials, in general there are no easily characterizable internal subsets of the parameter box Γ , whose Hurwitzness implies the Hurwitz invariance of the entire set [23]. In the same vein, it is shown in [16] that in considering the Hurwitz invariance of the characteristic polynomials of a polytope of $(n \times n)$ matrices, one has to check all $(2n - 4)$ -dimensional boundaries of the parameter space, a task which is computationally prohibitive even for matrices of reasonably small sizes. In such a case, the zero exclusion condition 2) above provides a relatively simple frequency sweeping procedure for verifying Hurwitz invariance. Such a graphical approach has been advocated through successful application in [9], [12], [13], and provides the conditions of [8], [9]. Recently, the so-called finite zero exclusion principle has been developed by Rantzer [21] to avoid frequency sweeping, and to thus permit fast computation. This approach too relies on the ready calculation of pertinent value sets. Besides their utility for robust stability analysis, value sets play an important role in the determination of the frequency response of a family of transfer functions; see, for example, [17]. Consequently, they can be used in designing robust controllers which meet performance considerations that go beyond mere closed-loop stabilization; see [22], for example.

Accordingly, in this paper, we consider both the determination of value sets as well as the Hurwitz invariance of (1.3). The principle contribution is to show that for the general family (1.7), at each fixed frequency ω , each member of $\partial H(\omega, \Gamma)$ has preimages in certain line segments in the parameter set Γ . These critical line segments are simply characterized, vary with frequency, are independent of the exponents k_i [see (1.7)], and consequently the Hurwitz invariance of the family of functions in (1.7) becomes equivalent to checking a finite number of continuous and piecewise differentiable scalar functions in ω for avoidance of the negative real axis.

For the case of (1.8), we show that the critical segments are, in fact, the edges of Γ plus certain simply constructable, frequency-independent, 45-degree line segments in the parameter space. Further, the Hurwitz invariance of $H(\Gamma)$ is guaranteed by that of these frequency-independent line segments (including the edges).

For the case of (1.9), we show that the critical lines determining the value set boundaries are either frequency independent or, as frequency varies, vary on certain (two-dimensional) planes and certain (three-dimensional) boxes in Γ . To check for robust Hurwitzness, it then suffices to check these frequency-independent lines, planes, and boxes.

It is instructive to compare this paper with the work of [24], which considers the set of uncertain polynomials

$$\sum \ddot{F}_i(s) \left(\prod \ddot{P}_{i,j}(s) \right) \quad (1.12)$$

where the $F_i(s)$ are fixed polynomials while the $P_{i,j}(s)$ vary in independent interval sets. Several points of difference are noteworthy. First, the set in (1.12) is broader than ours in the sense that it allows the sum of more than two factors. It is narrower than (1.7) in that the $P_{i,j}(s)$ vary in independent

intervals and have unity power, whereas in (1.7) arbitrary variations with arbitrary powers are allowed. However, the greatest difference lies in the approaches employed to obtain results derived in this paper and [24]. The latter employs a parameter space as opposed to the value set approach of Hurwitz invariance verification. Its result states that one has to check the Hurwitz invariance of internal manifolds to Γ of the dimension

$$\max \{n_i\}. \quad (1.13)$$

Thus, even for the cases of (1.8) and (1.9), [24] requires checking Hurwitz invariance over manifolds that have dimensions that increase with the number of factors in (1.8) and (1.9).

Section II considers the general case. The special cases of (1.8) and (1.9) are addressed in Sections III and IV, respectively. Section V is the conclusion.

II. VALUE SET BOUNDARIES

A major objective of this paper is to achieve the following. For a given frequency ω , identify a critical subset $\Gamma_c(\omega)$ of Γ having the property that for all

$$v \in \partial(H(\omega, \Gamma)) \quad (2.14)$$

there exists $\gamma \in \Gamma_c(\omega)$ such that

$$h(j\omega, \gamma) = v \quad (2.15)$$

i.e., every point on the boundary of the value set at this frequency has at least one preimage in $\Gamma_c(\omega)$. In characterizing such a $\Gamma_c(\omega)$, we will not attempt to extract the smallest possible such set, but will be content with one particular choice that enjoys the above properties, and at the same time has a relatively simple analytical description.

To this end, we adopt a somewhat indirect approach. Specifically, given ω , we give necessary conditions on a parameter vector γ such that

$$h(j\omega, \gamma) \in \partial(H(\omega, \Gamma)). \quad (2.16)$$

Clearly, parameter vectors satisfying these necessary conditions together suffice to define a $\Gamma_c(\omega)$ meeting the requirements specified above.

In the sequel, a k -side of the parameter box Γ will refer to a subset of Γ in which only k parameters vary and all others are fixed at their extreme values. By the same token, a point γ in the interior of a k -side has exactly k -elements that do not take extreme values. A point on the boundary of a k -side lies on this k -side, but not in its interior. If γ is in the interior of a k -side of Γ , then the elements of γ not fixed at extreme values will be called the variables on this k -side. We note, that every k -side of Γ is also a hyperrectangle, that all its l -sides, $l \leq k$ are also l -sides of Γ , and that corners and edges of Γ are its respective 0- and 1-sides. Further, the $\sum_{i=1}^n N_i$ -side of Γ (recall that this sum is the dimension of Γ) is Γ itself.

In characterizing $\Gamma_c(\omega)$, we will assume that all edges of Γ are automatically included in $\Gamma_c(\omega)$. Thereafter, for all

$$1 < k \leq \sum N_i \quad (2.17)$$

we will provide for each k -side of Γ necessary conditions for γ in its interior to obey (2.16).

The main results of this section can be summarized in the following way.

Result 1: Not all sides of Γ need contribute interior points to $\Gamma_c(\omega)$. In fact, certain sides are such that, irrespective of ω , their interior points need never be included in $\Gamma_c(\omega)$. In Section II-A, we give a result which characterizes what the sides that potentially contribute interior points to $\Gamma_c(\omega)$ are. A feature of this characterization is that, for each side which in Section II-A is identified as a potential interior point contributor, all the sides of this side are also similar contributors.

Result 2: Having eliminated a vast majority of sides by virtue of Result 1, we restrict attention to an arbitrary side Q of Γ whose interior points have been ascertained by Result 1 as being potential members of $\Gamma_c(\omega)$. For every such Q , at a given frequency ω , we associate a unique, possibly frequency-dependent affine line $L_a(Q)$ such that a γ in the interior of Q obeys (2.16) only if γ belongs to this line. Consequently, together with the edges of Γ , the union of the intersections of all $L_a(Q)$ with the interior of the corresponding contributing sides Q comprise $\Gamma_c(\omega)$. The affine line $L_a(Q)$ associated with a given contributing boundary Q is characterized in Section II-B.

The results to be presented will be illustrated through the following example.

$$h(j\omega, \gamma) = 1 + \frac{[(\omega^2 - 2)\gamma_{11} + (j\omega + 1)\gamma_{12}][\omega^2 + \gamma_{31} + \gamma_{32}]}{[(1 - \omega^2 + j\omega)\gamma_{21} + j\omega\gamma_{22}][0.5j\omega + \gamma_{41}]} \quad (2.18)$$

To proceed with the development, observe first that as far as the determination of $\Gamma_c(\omega)$ is concerned, one need only consider the transfer function

$$g(s, \gamma) = \prod_{i=1}^n (p_{i0}(s) + \gamma_i^T P_i(s))^{k_i} \quad (2.19)$$

For unless $g_0(j\omega)$ in (1.3) is zero, at any frequency ω , $g(j\omega, \gamma) \in \partial(G(\omega, \Gamma))$ (with $G(\omega, \Gamma)$, obviously defined), iff $h(j\omega, \gamma) \in \partial(H(\omega, \Gamma))$. Of course, if $g_0(j\omega) = 0$, then at this ω , $H(\omega, \Gamma)$ collapses to a single point, and any $\gamma \in \Gamma$, including any corner, describes $\partial(H(\omega, \Gamma))$. Thus, in this case, $\Gamma_c(\omega)$ can be trivially constituted by a solitary corner of Γ .

Thus, here onwards, attention will be restricted to the sets $g(j\omega, \gamma)$, $\partial(G(\omega, \Gamma))$ instead of $h(j\omega, \gamma)$, $\partial(H(\omega, \Gamma))$, respectively.

A. Result 1

We have the following theorem.

Theorem 2.1: For a given ω and every $v \in \partial(H(\omega, \Gamma))$, there exists a γ obeying (2.15), such that [see (1.4)], for each i , at most one γ_{i1} is a variable. Consequently, a boundary Q of Γ need contribute interior points to $\Gamma_c(\omega)$ only if, for each i , at most one γ_{i1} is a variable in the interior of Q .

Thus, for example, in (2.18), any side on which both γ_{11} , γ_{12} are variables is not included in $\Gamma_c(\omega)$. The proof of this theorem relies on the following lemma, proved in Appendix

A, which involves boundary determination of the power and the product of sets of complex numbers.

Lemma 2.1: Let $D_1, D_2, \dots, D_\sigma$ be bounded and closed sets of complex numbers. Define, for integers k_i ,

$$D_i^{(k_i)} := \{d_i^{k_i} : d_i \in D_i : i = 1, 2, \dots, \sigma\}, \quad (2.20)$$

and

$$\prod_{i=1}^{\sigma} D_i^{(k_i)} := \left\{ \prod_{i=1}^{\sigma} d_i^{k_i} : d_i \in D_i : i = 1, 2, \dots, \sigma \right\} \quad (2.21)$$

Then

$$\prod_{i=1}^{\sigma} (\partial D_i)^{(k_i)} \supset \partial \left(\prod_{i=1}^{\sigma} D_i^{(k_i)} \right) \quad (2.22)$$

We can now prove the theorem.

Proof of Theorem 2.1: In Lemma 2.1, identify

$$D_i := \{p_{i0}(j\omega) + \gamma_i^T P_i(s) : \gamma_i \in \Gamma_i\} \quad (2.23)$$

and $\sigma = n$. Then the result follows by noting that every element in $\partial(D_i)$ has at least one preimage in an edge of Γ_i [14].

B. Result 2

From here onwards, in determining contributions to $\Gamma_c(\omega)$ from the interior of a given side of Γ , attention need only be restricted to sides Q of Γ on which at most one element of each Γ_i is a variable.

Call such a prototype side Q . Lump the variables defining this side into the vector $q = [q_1, \dots, q_k]$, $k \leq n$. Suppose the coefficient polynomial of each q_i in (2.19) [i.e., the corresponding $P_{ij}(s)$], evaluated at $s = j\omega$, is nonzero. (If the coefficient polynomial is evaluated to be zero, then the corresponding q_i need not be included in q ; see discussion later.) Then, through a suitable extraction of the frequency-dependent coefficients of q_i in (2.19), at every ω , the image of the side Q under the mapping defined in (2.19), i.e., $G(\omega, Q)$, can be described by a set of complex numbers

$$f(Q) := \left\{ f(q) = f_0 \prod_{i=1}^k (q_i + \alpha_i + j\beta_i)^{k_i} : q \in Q \right\} \quad (2.24)$$

where f_0 is a complex constant, α_i and β_i , $i = 1, 2, \dots, k$ are real constants, and $k_i \neq 0$ for all $i = 1, 2, \dots, k$.

Fact 1: Notice that, as ω varies, f_0 , α_i , and β_i vary in a rational fashion with ω . Further, $f(Q)$ is bounded because of Assumption 1.3 and the fact that the coefficient polynomial of each q_i in (2.19) evaluated at $s = j\omega$ is nonzero.

Thus, for example, in (2.18), consider the interior of the 2-side defined by $\gamma_{12} = \gamma_{12}^+$, $\gamma_{32} = \gamma_{32}^-$, $\gamma_{21} = \gamma_{21}^+$, $\gamma_{41} = \gamma_{41}^+$, $\gamma_{22} = \gamma_{22}^-$. In this case, $q_1 = \gamma_{11}$, $q_2 = \gamma_{31}$:

$$Q := \{q = [q_1, q_2]^T : q_i^- \leq q_i \leq q_i^+, i = 1, 2\}. \quad (2.25)$$

Further

$$f(q) = \frac{(\omega^2 - 2)}{[(1 - \omega^2 + j\omega)\gamma_{21}^+ + j\omega\gamma_{22}^-][0.5j\omega + \gamma_{41}^+]} \cdot \left[q_1 + \frac{(j\omega + 1)\gamma_{12}^+}{(\omega^2 - 2)} \right] [q_2 + \omega^2 + \gamma_{32}^-]. \quad (2.26)$$

Observe that, at $\omega^2 = 2$, the coefficient of γ_{11} ($= q_1$) is zero. Consequently, the representation in (2.26) is infeasible. However, we argue now that at this ω , no interior point of Q under consideration need be included in $\Gamma_c(\omega)$. This is because, at this ω , $H(\omega, \Gamma)$ is independent of $\gamma_{11} = q_1$. Thus, if we select γ_{11} at an extreme value, without changing the variable $\gamma_{31} = q_2$, one does not alter the value of $h(j\omega, \gamma)$. Thus, corresponding to each point q in the interior of Q , there lies a point q^* on a boundary of Q having precisely the same image in the value set space as does q . Then, in determining contributors to $\Gamma_c(\omega)$, one need not consider any point in the interior of Q , as these points are covered by points on the boundary of Q . This observation leads to the following formal fact.

Fact 2: Suppose Q is a side of Γ that conforms to the requirements of Theorem 2.1. Further suppose, with q_i the variables on Q , at some frequency ω , and some j , the coefficient polynomial of q_j in (1.3) is zero. Then, $\Gamma_c(\omega)$ will not contain any points in the interior of Q . Also, if the coefficient polynomial of each q_i is nonzero, then f_0 in (2.24) is nonzero.

Finally, observe, from the foregoing that the basis for limiting the sides that contribute to $\Gamma_c(\omega)$ is that the edges of each individual factor by themselves cover the value set boundary of that factor. Although, in general, an N -dimensional polytope has $N2^{N-1}$ edges, at a given frequency, at most $2N$ of these edges need be considered for constructing the value set boundary of the polytope. These special $2N$ edges are easily characterized (see [18]). Thus, the number of contributing sides is even smaller than that specified in Theorem 2.1. However, to prevent notational encumbrances, henceforth we will adhere to the somewhat conservative characterization given in Theorem 2.1.

Provided in (2.24), $\beta_i \neq 0, \forall i \in \{1, \dots, k\}$, we will call the affine line in (2.27), below, the line associated with Q

$$L_a(Q) := \{(q_1 q_2 \dots q_k)^T = \rho(\beta_1 \beta_2 \dots \beta_k)^T : \rho(\alpha_1 \alpha_2 \dots \alpha_k)^T : -\infty < \rho < \infty\}. \quad (2.27)$$

The situation where one or more of the β_i equal zero will be dealt with later. Observe that the intersection of $L_a(Q)$ with the interior of Q is given by the segment

$$L(Q) = \{q = (q_1 \dots q_k)^T : q_i = \rho\beta_i - \alpha_i, \forall i \in \{1, \dots, k\}, \rho^-(Q) < \rho < \rho^+(Q), q \in Q\} \quad (2.28)$$

where

$$\rho^-(Q) = \max_{i \in \{1, \dots, k\}} \min \left\{ \frac{q_i^+ + \alpha_i}{\beta_i}, \frac{q_i^- + \alpha_i}{\beta_i} \right\} \quad (2.29)$$

and

$$\rho^+(Q) = \min_{i \in \{1, \dots, k\}} \max \left\{ \frac{q_i^+ + \alpha_i}{\beta_i}, \frac{q_i^- + \alpha_i}{\beta_i} \right\} \quad (2.30)$$

where q_i^+ and q_i^- are the extreme values of the variable q_i . Note the following important facts.

Fact 3: If $\rho^+(Q) \leq \rho^-(Q)$, then this set is empty, in which case, as will be shown soon, the interior of Q contributes nothing to $\Gamma_c(\omega)$.

Fact 4: As the α_i, β_i depend on ω (Fact 1), the line $L(Q)$. Equally, while at some frequencies, $L(Q)$ may be empty, at others, it may well be.

Fact 5: As noted earlier, whenever Q conforms to the requirements of Theorem 2.1, so do all its boundaries. It is easy to see from the foregoing definitions that the line associated with a boundary of Q is, in fact, a projection of the above line onto this boundary.

We now present the main result of this section, proved in Appendix B.

Theorem 2.2: Consider Q a boundary of Γ obeying the conditions of Theorem 2.1, with the various quantities as defined in the foregoing and $f_0 \neq 0$ (see Fact 2).

i) Suppose in (2.24) that, for some $i \in \{1, \dots, k\}, \beta_i = 0$. Then for every q in the interior of Q such that $f(q) \in \partial(f(Q))$, there exists a q^* on an edge of Q such that $f(q) = f(q^*)$. ii) Suppose $\beta_i \neq 0, \forall i \in \{1, \dots, k\}$, and there exists q in the interior of Q such that $f(q) \in \partial(f(Q))$. Then $q \in L(Q)$. iii) Suppose, in addition to the conditions set out in ii)

$$\sum_{i=1}^n k_i = 0. \quad (2.31)$$

Then there exists q^* on a boundary of Q such that for all $q \in L(Q)$

$$f(q) = f(q^*). \quad (2.32)$$

Before illustrating this theorem with the example of (2.18), we highlight some of its features. First of all, if any $\beta_i = 0$, then no point in the interior of Q is included in the formation of $\Gamma_c(\omega)$. For, given any γ in the interior of Q obeying (2.16), there is a parameter vector on an edge of Γ that has the same image in the value set space as does γ . Since $\Gamma_c(\omega)$ already includes the edges of Γ , the image of this γ stands covered from the outset.

Likewise, if (2.31) holds, then all interior points of Q mapping to the value set boundary have the same image on this boundary, an image also shared by another parameter vector lying in the interior of a boundary of Q . Thus, these points need not be included as they will be covered when one considers the boundaries of Q .

Further, in the event that i), iii) of the theorem do not hold, then the critical subset contributed by interior points of Q lies exclusively on the associated line segment. Thus, indeed, at every ω , $\Gamma_c(\omega)$ comprised exclusively of line segments, at most one for each side conforming to the requirements of Theorem 2.1. Of course, these segments in general vary with ω , and should their intersection with an associated Q be empty, then the interior of that Q fails to contribute elements to $\Gamma_c(\omega)$.

Finally, we note that the equations describing the internal line segments do not depend on the powers μ_i and ν_i .

We now illustrate these results by invoking the example of (2.18). In the example, assume $\omega = 2$, $\gamma_{1j}^- = 0$, and $\gamma_{1j}^+ = 1$.

Consider first any side on which γ_{31} or γ_{32} are variables [recall that if both were variables, then such a side will not contribute an interior point to $\Gamma_c(\omega)$]. Then i) of Theorem 2.2 holds, and no point in the interior of this boundary is included in $\Gamma_c(\omega)$.

Next, consider the 2-side defined by $\gamma_{31} = \gamma_{41} = 0$ and $\gamma_{32} = \gamma_{12} = \gamma_{22} = 1$. Then, on this boundary, with $q_1 = \gamma_{11}$ and $q_2 = \gamma_{21}$, (2.31) holds. Consequently, this side too contributes no interior points to $\Gamma_r(\omega)$.

Finally, consider the 3-side $\gamma_{12} = \gamma_{21} = 1$ and $\gamma_{31} = \gamma_{32} = 0$. Choose $q_1 = \gamma_{11}$, $q_2 = \gamma_{22}$, and $q_3 = \gamma_{41}$. Then, we have

$$f(q) = 4j \frac{q_1 + 0.5 + j}{[q_2 + 1 + 1.5j][q_3 + j]}. \quad (2.33)$$

Then the interior points of such a 3-side to be included in $\Gamma_r(\omega)$ are given by the segment

$$L(Q) = \{[q_1, q_2, q_3] = \rho[1, 1.5, 1] - [0.5, 1, 0] : 2/3 < \rho < 1\}. \quad (2.34)$$

C. Hurwitz Invariance

Having shown that $\partial(H(j\omega, \Gamma))$ is mapped from the critical line segments in Γ , we now turn to verifying the Hurwitz invariance of (1.7). Essentially, $\partial(H(j\omega, \Gamma))$ must be checked for zero exclusion.

In view of the definition of $\Gamma_r(\omega)$, the Hurwitz invariance of (1.7) is equivalent to the requirement that: i) at least one member of (1.7) is Hurwitz, and ii) that at every ω and all $\gamma \in \Gamma_r(\omega)$

$$h(j\omega, \gamma) \neq 0. \quad (2.35)$$

Since at every ω , $\Gamma_r(\omega)$ comprises the edges of Γ and certain ω -specific line segments, to check for Hurwitz invariance, it suffices to check that all transfer functions corresponding to the edges of (1.7) are Hurwitz, and that the image, in the value set space, of each aforementioned ω -specific line segment is zero exclusive. The principal contribution of this section is to demonstrate that the zero exclusivity of these segments can be verified by checking certain piecewise continuous and piecewise differentiable functions of ω for avoidance of the negative real axis.

To avoid notational complexities, we show this fact in a somewhat informal fashion. Consider a side Q of Γ , which meets the requirements implicit in Theorem 2.1. Recall that for each such Q , potentially there exists an internal line segment that contributes to $\Gamma_r(\omega)$.

For a prototype Q , the internal line segment is as given in (2.28)–(2.30), with the value set $G(\omega, \Gamma)$ as in (2.24). Observe that the various quantities in these equations depend on ω , and that at certain frequencies, this segment could be empty (Facts 3 and 4), or may not be a member of $\Gamma_r(\omega)$ (e.g., when i) and/or iii) of Theorem 2.2 holds, or when f_0 (see (2.24) and Fact 2) or $g_0(j\omega)$ is zero. Call the set of frequencies at which $L(Q)$ is a nonempty subset of $\Gamma_r(\omega)$, $\Omega(Q)$. Then we need to check that the image in the value set space of $L(Q)$ is zero exclusive at all $\omega \in \Omega(Q)$. Direct substitution of (2.28)–(2.30) into (2.24), together with the relation among $G(\omega, \Gamma)$, $H(\omega, \Gamma)$, and (2.24), shows that at all $\omega \in \Omega(Q)$, the image of $L(Q)$ is given by (2.36)–(2.38). In these equations, to avoid confusion, unlike (2.28)–(2.30) and (2.24) the various quantities have been expressed explicitly as functions of ω

and Q

$$h(j\omega, \gamma) = 1 + F(Q, \omega)(j + \rho)^{M(Q)}; \quad \rho^-(Q, \omega) < \rho < \rho^+(Q, \omega) \quad (2.36)$$

where

$$F(Q, \omega) = g_0(j\omega)f_0(Q, \omega) \prod_{i=1}^{k(Q)} \beta_i(Q, \omega)^{k_i(Q)} \quad (2.37)$$

and

$$M(Q) = \sum_{i=1}^{k(Q)} k_i(Q). \quad (2.38)$$

Recall that the choice of Q (see (2.31) and the discussion subsequent to Theorem 2.2) ensures that $M(Q) \neq 0$. Moreover, from the definition of $\Omega(Q)$,

$$F(Q, \omega) \neq 0, \quad \forall \omega \in \Omega(Q) \quad (2.39)$$

and

$$\rho^-(Q, \omega) < \rho^+(Q, \omega), \quad \forall \omega \in \Omega(Q). \quad (2.40)$$

We claim that the bounds $\rho^+(Q, \omega)$ and $\rho^-(Q, \omega)$ are continuous and piecewise differentiable in these frequency ranges. To see this, we note that $\alpha_i(Q, \omega)$ and $\beta_i(Q, \omega)$ [see (2.24)] are continuous and differentiable. The minmax functions in (2.29)–(2.30) preserve continuity, and the lack of differentiability occurs at the isolated frequencies where the minmax selections “change.”

Thus, from (2.36), one can see that (2.35) holds for all γ in the interior of this prototype Q and belonging to $\Gamma_r(\omega)$ iff

$$\mu(Q, \omega) := (-F(Q, \omega))^{-1/M(Q)} - j \notin (\rho^-(Q, \omega), \rho^+(Q, \omega)), \quad \forall \omega \in \Omega(Q). \quad (2.41)$$

In the above, if $M(Q) \neq \pm 1$, then all the roots of $(-F(Q, \omega))^{-1/M(Q)}$ should be considered. Define the functions

$$\xi(Q, \omega) = \begin{cases} \frac{\mu(Q, \omega) - \rho^-(Q, \omega)}{\mu(Q, \omega) - \rho^+(Q, \omega)} & \forall \omega \in \Omega(Q) \\ 1 & \text{otherwise.} \end{cases} \quad (2.42)$$

The functions $\xi(Q, \omega)$ defined above are piecewise continuous and differentiable. Moreover, $\forall \omega \in \{(-\infty, \infty) \cap \Omega(Q)\}$, $\xi(Q, \omega) = 1$. Recall that, at these frequencies, there are no contributions from the interior of Q to $\Gamma_r(\omega)$, and (2.41) need not be checked. Furthermore, the required zero exclusion of $h(j\omega, \gamma)$ for all γ in the interior of this prototype Q and belonging to $\Gamma_r(\omega)$ is equivalent to

$$\xi(Q, \omega) \notin (-\infty, 0). \quad (2.43)$$

We therefore have the following theorem.

Theorem 2.3: The family of transfer functions $H(\Gamma)$ described in (1.7) is Hurwitz invariant if and only if the following conditions hold:

a) $h(s, \gamma)$ is Hurwitz for all γ in the edges of Γ ; b) For each $\omega \in \mathbb{R}$, and all Q satisfying the requirements of Theorem 2.1, the piecewise continuous and differentiable functions $\xi(Q, \omega)$ defined in (2.42) avoid the negative real axis.

III. REAL ZERO-POLE-GAIN VARIATIONS

We now consider the special case of (1.2) where the certain parameters are real poles, zeros, and gains.

Consider the family of transfer function $H(\Gamma)$ described by (1.8) and (1.2). The parameters d and λ_i vary independently within given bounds, i.e.,

$$d^- \leq d \leq d^+; \lambda_i^- \leq \lambda_i \leq \lambda_i^+, \quad i = 1, 2, \dots, n-1 \quad (3.44)$$

and

$$\gamma \in \Gamma = [\lambda_1^-, \lambda_1^+] \times \dots \times [\lambda_{n-1}^-, \lambda_{n-1}^+] \times [d^-, d^+]. \quad (3.45)$$

Notice that Assumption 1.3 implies that if a given k_i is negative, the corresponding interval of λ_i cannot include zero.

Recall, from Section II, that apart from the edges of Γ , we need to consider an internal segment associated with each side of Γ , which conforms to the requirements of Theorem 2.1. Observe that every k -side Q of Γ , $k \geq 1$ conforms to this requirement. Consider, now, two possible cases of such k -sides Q .

Case I: The interior of Q has d as a variable.

Observe that, with $q_1 = d$ in (2.24), $\beta_1 = 0$. Thus, because of i) of Theorem 2.2, no parameter vector in the interior of such a Q is included in $\Gamma_c(\omega)$. Consequently, for each $\gamma \in \Gamma_c(\omega)$, d is at an extreme value.

Case II: The variables of Q exclude d

Assume the variables in Q are for some $S \subset \{1, \dots, n-1\}$, λ_i , $i \in S$. Observe, in (2.24), that each $\beta_i = \omega$ and $\alpha_i = 0$. From i) of Theorem 2.2, one concludes that at $\omega = 0$, $\Gamma_c(\omega)$ comprises only the edges of Γ . Also from Theorem 2.2, when $\omega \neq 0$, a parameter vector in the interior of Q is in $\Gamma_c(\omega)$ only if it obeys

$$\lambda_i = \rho\omega, \forall \rho^-(Q) < \rho < \rho^+(Q), \quad i \in S \quad (3.46)$$

where

$$\rho^-(Q) = \max_{i \in S} \lambda_i^- \quad (3.47)$$

and

$$\rho^+(Q) = \min_{i \in S} \lambda_i^+ \quad (3.48)$$

In addition [ii) of Theorem 2.2], we must have

$$\sum_{i \in S} k_i \neq 0. \quad (3.49)$$

Thus, we have the following theorem.

Theorem 3.1: Consider the parameter box Γ in \mathbb{R}^n given in (3.45). Then $\Gamma_c(\omega)$ is frequency invariant, and comprises the edges of Γ , and all line segments of the form in (3.51)–(3.53), for every $S \subset \{1, \dots, n-1\}$, obeying (3.49) and

$$\max_{i \in S} \lambda_i^- \leq \lambda_j \leq \min_{i \in S} \lambda_i^+, \quad \forall j \in S. \quad (3.50)$$

$$\lambda_i = \lambda_j, \quad \forall i, j \in S \quad (3.51)$$

$$\lambda_j \in \{\lambda_j^+, \lambda_j^-\}, \quad \forall j \notin S \quad (3.52)$$

$$d \in \{d^+, d^-\}.$$

Furthermore, the family of functions $H(\Gamma)$ is Hurwitz invariant if and only if $h(s, \gamma)$ is Hurwitz invariant on the 45-degree line segments and the edges of Γ .

A. A Special Case

Notice that if some of the λ_i vary in nonoverlapping intervals, then the number of critical segments reduces, as can be easily seen from (3.51). A case of special interest is when all the poles and zeros vary in intervals that do not overlap. In such an event, the projections mentioned above are empty, and edges suffice for value set boundary and Hurwitz invariance. One thus obtains the corollary below, itself a variation of the results (see Remark 3.3 for comparison) of [25], [26].

Corollary 3.1 (An Edge Theorem): Consider the parameter box Γ in (3.45) and the family of function $H(\Gamma)$ described by (1.8). Suppose $(\lambda_i^-, \lambda_i^+) \cap (\lambda_j^-, \lambda_j^+)$ are empty for all $1 \leq i < j < n$. Then the boundary of the value set $H(\omega, \Gamma)$ at any frequency ω is mapped from the edges of Γ . Furthermore, $H(\Gamma)$ is Hurwitz invariant if and only if all the edges of $H(\Gamma)$ are Hurwitz invariant.

Remark 3.1: For overlapping intervals of λ_i , however, the 45-degree line segments are indeed necessary. To show this, we provide the following simple example. Consider

$$h(s, \lambda_1, \lambda_2) = \frac{0.1(0.8s^2 + 0.8s + 4.5)(s + \lambda_1)(s + \lambda_2)}{s^4 + 10s^3 + 11.8s^2 + 11.8s + 0.2} \quad (3.54)$$

$$\lambda_1, \lambda_2 \in [-30, 0].$$

In this example, there are four edges with the associated transfer functions given by

$$h(s, 0, \lambda_2), \lambda_2 \in [-30, 0],$$

$$h(s, -30, \lambda_2), \lambda_2 \in [-30, 0],$$

$$h(s, \lambda_1, 0), \lambda_1 \in [-30, 0];$$

$$h(s, \lambda_1, -30), \lambda_1 \in [-30, 0].$$

There is only one 45-degree line segment given by

$$h(s, \lambda, \lambda), \lambda \in [-30, 0].$$

It is straightforward to verify that the transfer functions on all the edges are Hurwitz, but some on the 45-degree line segment are not. For example, at $\lambda = -15$, $h(s, \lambda, \lambda)$ has the unstable zeros $0.2424 \pm 1.8914j$.

Remark 3.2: In actual fact, the result in [26] is stronger. It requires that only $2n$ edges be considered. This fact is also recoverable from [27], which also employs the Jacobian rank deficiency approach underlying our development.

Remark 3.3: Neither [26] nor [25] deals with the overlapping root situation considered in Theorem 3.1 above; nor do [26] and [25] permit the structural preservation of pole-zero multiplicities.

IV. COMPLEX ZERO-POLE-GAIN VARIATIONS

In this section, the complex zero-pole-gain variations case of (1.9) is treated. As before, $\Gamma_c(\omega)$ comprise line segments.

We note that the case $\omega = 0$ is trivial. At this frequency, for every Q obeying Theorem 2.1, all the β_i in (2.24) are zero. Thus, from Theorem 2.2, $\Gamma_c(0)$ comprises the edges of Γ only, and $\omega \neq 0$ will be assumed in the sequel.

The characterization of the line segments follows as in Theorems 2.1 and 2.2. In this section, we will focus mainly on showing that these frequency-dependent line segments obey the confinement rules stated in the Introduction. To this end, we consider the various possible Q obeying Theorem 2.1, and consider how the associated line segments change with frequency. Observe, as in iii) of Theorem 2.2, that certain sides Q can be eliminated according to the combinations of powers of their defining factors. In the sequel, we will only consider Q on which (2.31) does not hold. Then the only restriction that Theorem 2.2 places on Q is that for no i can both a_i and b_i simultaneously be variables. Consider, now, the following possible cases of Q .

Case 1: d is a variable on Q . As in Section III, such a Q does not contribute interior points to $\Gamma_c(\omega)$.

Here onwards, we assume d is at an extreme value. In the sequel, without loss of generality, all nonvariables will be denoted with a superscript " \pm ," and the λ_i will always be considered as potential variables.

Case 2: No a_i nor any b_i is a variable. As in Section III, an interior point of Q is in $\Gamma_c(\omega)$ only if it is on a frequency-invariant 45-degree line segment similar to that in Section III. The set of all such segments will be called L_1 .

Case 3: No a_i is a variable, but some b_i are. Suppose, as before, that q_i are the variables. Then if $q_i = \lambda_i$, in the corresponding factor in (2.24), $\alpha_i = 0$ and

$$\beta_i = \omega. \quad (4.55)$$

Further, if $q_i = b_i$, then in the corresponding factor in (2.24)

$$\alpha_i = -\omega^2 \quad (4.56)$$

and

$$\beta_i = \omega a_i^\pm. \quad (4.57)$$

Thus, on the line segment associated with Q , the variable λ_i and b_i obey

$$\lambda_i = 0 + \rho\omega \quad (4.58)$$

$$b_i = -\omega^2 + \rho\omega a_i^\pm. \quad (4.59)$$

Defining $\rho_1 = \rho\omega$ and $\rho_2 = \omega^2$, one finds that all interior points of Q to be included in $\Gamma_c(\omega)$ lie on the two-dimensional plane,

$$\gamma = \rho_1 C_1 + \rho_2 C_2 \quad (4.60)$$

where $C_1, C_2 \in \mathbb{R}^N$ are frequency-independent constant vectors.

Case 4: No b_i is a variable. If $q_i = a_i$, then in the corresponding factor in (2.24)

$$\alpha_i = 0 \quad (4.61)$$

and

$$\beta_i = \omega - \omega^{-1} b_i^\pm \quad (4.62)$$

Then, as long as all the $\beta_i \neq 0$, on the line segment associated with Q , the variable λ_i obey (4.58) and a_i obey

$$a_i = 0 + \rho(\omega - \omega^{-1} b_i^\pm) \quad (4.63)$$

Thus, again, this segment is confined to a plane of the form of (4.60) with different C_1 and C_2 . The corresponding ρ_1 and ρ_2 in this case should be defined by $\rho_1 = \rho\omega$ and $\rho_2 = \rho\omega^{-1}$.

Case 5: For some i , a_i is a variable, for some others, b_i is a variable. As before, on the line segment associated with Q , the variable λ_i , b_i , and a_i , respectively, obey (4.58), (4.59), and (4.63). Then with $\rho_1 = \rho\omega$, $\rho_2 = \omega^2$, $\rho_3 = \rho\omega^{-1}$, and C_i , $i = 1, 2, 3$ frequency-independent constant vectors (C_1, C_2 different from those in Case 3), this segment can be seen to lie on a three-dimensional plane of following form

$$\gamma = \rho_1 C_1 + \rho_2 C_2 + \rho_3 C_3 \quad (4.64)$$

The set of two-dimensional planes covered by Cases 3 and 4 will be denoted L_2 , while the set of boxes in Case 5 will be called L_3 . The zero exclusion principle then immediately yields the following theorem.

Theorem 4.1: With L_1 , L_2 , and L_3 defined as above, the family of transfer functions $H(\Gamma)$ described by (1.9) is Hurwitz invariant if and only if all the edges, the internal segments in L_1 , the rectangles in L_2 , and the boxes in L_3 are Hurwitz invariant.

We note that the sets in L_1 , L_2 , and L_3 are easily characterized from the critical frequency-dependent segments they contain. As in Section III, many of these segments, rectangles, and boxes will be empty.

V. CONCLUSIONS

In this paper, we have considered robust stability verification of linear time-invariant systems characterized by the class of nonlinear parametric perturbations given in (1.7). In light of the zero exclusion principle, our focus has been on both the verification of Hurwitz invariance and the construction of value sets for the system characteristic function. The main result on construction of value sets shows that for the class of nonlinear parametric perturbations given in (1.7), the value set boundary of the characteristic function at each fixed frequency is determined by the edges and some frequency-dependent internal line segments in the parameter box. This result greatly simplifies the construction of the value sets, and considerably eases the task of robust stability verification. Indeed, a piecewise continuous and differentiable frequency sweeping function is found such that Hurwitz invariance of the set in question is equivalent to this function's avoidance of the negative real axis. For the special case of real zero-pole gain variations, the critical line segments are all frequency independent; hence, the determination of robust

stability is even simpler. For the case of complex zero-pole variations, the critical internal lines are either frequency dependent or vary in certain (two-dimensional) planes or (three-dimensional) boxes.

The key device used in our development concerns a Jacobian function which helps isolate certain critical subsets of the parameter box whose elements collectively determine the value set boundary. This Jacobian-based technique may provide an effective tool for the robust stability analysis of sets which are even more general than the ones considered here. Indeed, a similar device is featured in [7], [28], [27] in relation to the multilinear problem.

APPENDIX A

PROOF OF LEMMA 2.1

We prove this result in two parts. First, we show that for D a bounded and closed set of complex numbers,

$$(\partial D)^{(k)} \supset \partial(D^{(k)}). \quad (\text{A.65})$$

We then show that for $D_1, D_2, \dots, D_\sigma$ bounded and closed sets of complex numbers, with

$$\begin{aligned} \prod_{i=1}^{\sigma} D_i &:= \left\{ \prod_{i=1}^{\sigma} d_i : d_i \in D_i : i = 1, 2, \dots, \sigma \right\} \\ \prod_{i=1}^{\sigma} (\partial D_i) &\supset \partial \left(\prod_{i=1}^{\sigma} D_i \right). \end{aligned} \quad (\text{A.66})$$

Together, these two parts prove the result.

Proof of Part 1: Given any complex number $z \in \partial(D^{(k)})$, we need to show that $z \in (\partial D)^{(k)}$. By the boundedness of $D^{(k)}$ (from that of D), there exists a sequence of complex numbers $\{z_j\}$ outside $D^{(k)}$ such that $z_j \rightarrow z$ as $j \rightarrow \infty$. Since $D^{(k)}$ is closed (as D is closed), there exists some $d \in D$ such that $d^k = z$. Define $d_{1j}, d_{2j}, \dots, d_{kj}$ to be the k th roots of z_j . Then for all i, j , $d_{ij} \notin D$ because $z_j = (d_{ij})^k \notin D^{(k)}$. On the other hand, the sequence $\{d_{ij}\}$ has a subsequence converging to $d \in D$. It follows that $d \in \partial D$, or equivalently, $z = d^k \in (\partial D)^{(k)}$.

Proof of Part 2: Given any complex number $z \in \partial(\prod_{i=1}^{\sigma} D_i)$, we need to show $z \in \prod_{i=1}^{\sigma} (\partial D_i)$. This clearly holds if one of the D_i is just $\{0\}$. So, assume every D_i contains at least one nonzero element. This implies that every ∂D_i has at least one nonzero element. By the boundedness of $(\prod_{i=1}^{\sigma} D_i)$ (from that of D_i), there exists a sequence of complex numbers $\{z_j\}$ outside $(\prod_{i=1}^{\sigma} D_i)$ such that $z_j \rightarrow z$ as $j \rightarrow \infty$. Since $(\prod_{i=1}^{\sigma} D_i)$ is closed (as D_i is closed), there exist some $d_i \in D_i$, $i = 1, 2, \dots, \sigma$ such that $z = d_1 d_2 \cdots d_\sigma$. If z is zero, it follows that all d_i , except for one which is set to zero, can be chosen to be a nonzero boundary point of ∂D_i . In the sequel, we assume that at most one d_i is zero. We claim that $d_i \in \partial D_i$, $i = 1, 2, \dots, \sigma$. Without loss of generality, consider d_1 . From the foregoing, one can choose $d_1 \in \partial D_1$ if one of the other d_i is zero. On the other hand, if the remaining d_i are nonzero, we define

$$d_{1j} := d_2 \cdots d_\sigma \quad j = 1, 2, \dots \quad (\text{A.67})$$

Then, d_{1j} converge to d_1 as $j \rightarrow \infty$ because d_i are bounded. Yet $d_{1j} \notin D_1$ for all j because $z_j = d_1 d_2 \cdots d_\sigma \notin D$. Therefore, $d_1 \in \partial D_1$. Similarly, it can be shown that $d_i \in \partial D_i$ for all other i ; hence, the result holds.

APPENDIX B

PROOF OF THEOREM 2.2

To prove the theorem, we need three lemmas. The first of these is well known (see, e.g., [28], [27]), and hence its proof is omitted.

Lemma B.1: Consider the hyperrectangle

$$Q := \{q = (q_1 q_2 \cdots q_m)^T : q_i^- \leq q_i \leq q_i^+, i = 1, 2, \dots, m\} \subset \mathbb{R}^m \quad (\text{B.68})$$

and a differentiable complex valued function $f(\cdot) : Q \rightarrow \mathbb{C}$ with all its first derivatives continuous. In the sequel, we will denote

$$f(Q) := \{f(q) : q \in Q\}. \quad (\text{B.69})$$

Then a point q in the interior of Q obeys $f(q) \in \partial f(Q)$ only if the following Jacobian matrix, evaluated at q , has rank less than two

$$J_f(q) := \begin{bmatrix} \operatorname{Re} \left(\frac{\partial f(q)}{\partial q_1} \right) & \operatorname{Re} \left(\frac{\partial f(q)}{\partial q_2} \right) & \operatorname{Re} \left(\frac{\partial f(q)}{\partial q_m} \right) \\ \operatorname{Im} \left(\frac{\partial f(q)}{\partial q_1} \right) & \operatorname{Im} \left(\frac{\partial f(q)}{\partial q_2} \right) & \operatorname{Im} \left(\frac{\partial f(q)}{\partial q_m} \right) \end{bmatrix} \quad (\text{B.70})$$

It should be noted that the lemma above is of little value when $f(Q)$ degenerates to a real segment. For when $f(\cdot)$ is real, the second row of the matrix in (B.70) is identically zero, and hence rank deficiency occurs for all q . For such a case, we have the following result.

Lemma B.2. Consider the hyperrectangle Q in (B.68) and a real continuous function $f(\cdot) : Q \rightarrow \mathbb{R}$. Then $f(Q)$ can be mapped from the edges of Q if and only if the two extreme points (minimum and maximum) of $f(Q)$ can be mapped from the edges of Q .

Proof: Necessity is obvious. To prove sufficiency, suppose q^1 and q^2 are the edge points corresponding to the extreme points of $f(Q)$. Observe that q^1 and q^2 can be connected by a path entirely in the edges of Q . By continuity of $f(\cdot)$, the image of this path, which is a subset of the edges, covers the whole of $f(Q)$.

The final lemma needed is given below.

Lemma B.3: Consider the hyperrectangle Q and the bounded set $f(Q)$ in (2.24) with $\beta_i = 0, \forall i$. Then each point in $f(Q)$ has at least one preimage in the edges of Q .

Proof: Observe that the result will not be affected by the value of f_0 . So choose $f_0 = 1$. Using Lemma B.2, we simply need to show that both the minimum and the maximum of $f(Q)$ have preimages in the edges of Q . Take the minimum, for example, as the maximum can be dealt with in the same manner. Denote the minimum by f_m and consider the two cases: 1) $f_m = 0$; and 2) $f_m \neq 0$. Case 1) implies that some $q_i + \alpha_i$ is zero with $k_i > 0$. In this case, it is obvious that setting the other q_i at extreme values does not change f_m . In Case 2), we claim that all the q_i must take their extreme value. Indeed, if some q_i were not at its extreme, f_m would not be

minimum because we can decrease the value of the function by increasing or decreasing this q_i . Therefore, in both cases, f_m can be achieved at an edge point.

Proof of Theorem 2.2: Consider a point q in the interior of Q such that

$$f(q) \in \partial f(Q). \quad (\text{B.71})$$

Proof of i): Suppose first that $\beta_i = 0, \forall i$. Then, Lemma B.3 proves the conclusions of i). Next, suppose that at least one β , without loss of generality β_1 , is nonzero. Suppose also, that for some q in the interior of Q , $f(q) \in \partial f(Q)$. Then from Lemma B.1, $\forall i, j \in \{1, \dots, k\}$, there exist real scalars x and y , not both zero, such that

$$x \frac{\partial f(q)}{\partial q_i} = y \frac{\partial f(q)}{\partial q_j}. \quad (\text{B.72})$$

Thus, from (2.24), the above gives

$$\frac{k_i f(q)}{q_i + \alpha_i + j\beta_i} = \frac{k_j f(q)}{q_j + \alpha_j + j\beta_j} \quad (\text{B.73})$$

which simplifies to

$$(q_i + \alpha_i)\beta_j = (q_j + \alpha_j)\beta_i. \quad (\text{B.74})$$

Now, suppose that at least one $\beta_i, i \neq 1$, without loss of generality β_2 , equals zero. Then with $i = 1, j = 2$, one has from (B.74) that

$$q_2 + \alpha_2 = 0. \quad (\text{B.75})$$

Thus, $f(q) = 0$. Moreover, this holds no matter what value the $q_i, i \neq 2$ take. Setting these $q_i, i \neq 2$ to their respective extreme values, one proves i).

Proof of ii): Follows from (B.74).

Proof of iii): Direct substitution of (B.74) into (2.24) yields

$$f(q) = c(\rho + j)^M \quad (\text{B.76})$$

where c is a suitable complex constant and the integer M is given by

$$M = \sum_{i=1}^k k_i. \quad (\text{B.77})$$

Thus, when $M = 0$, $f(q)$ has the same value on the whole segment $L(Q)$. Thus, by continuity, the image of this entire segment is covered by any one of its endpoints which is on a boundary of Q .

ACKNOWLEDGMENT

The authors are grateful to the members of Laboratoire d'Automatique at UCL for having fostered a productive, intellectually stimulating, yet relaxed atmosphere which is remarkably conducive to research.

REFERENCES

- [1] A. Vicino, A. Tesi, and M. Milanese, "An algorithm for nonconservative stability bounds computation for systems with nonlinearly correlated parametric uncertainties," in *Proc. 27th Conf. Decision Contr.*, Austin, TX, vol. 3, 1988, pp. 1761-1766.
- [2] V. Balakrishnan, S. Boyd, and S. Balemi, "Branch and bound algorithm for computing the minimum stability degree of parameter-dependent linear systems," Tech. Rep., Inform. Syst. Lab., Stanford Univ., 1991.
- [3] V. L. Kharitonov, "Asymptotic stability of an equilibrium position of a family of systems of linear differential equations," *Differential Equations*, vol. 14, pp. 1483-1485, 1979.
- [4] A. C. Bartlett, C. V. Hollot, and L. Huang, "Root locations of an entire polytope of polynomials: It suffices to check the edges," *Math. Contr., Signals, Syst.*, vol. 1, pp. 61-71, 1989.
- [5] M. Fu and B. R. Barmish, "Polytopes of polynomials with zeros in a prescribed set," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 544-546, May 1989.
- [6] H. Chapellat and S. P. Bhattacharyya, "A generalization of Kharitonov's theorem for robust stability of interval plant," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 306-312, 1989.
- [7] F. J. Kraus, M. Mansour, and B. D. O. Anderson, "Robust stability of polynomials with multilinear parametric dependence," *Int. J. Contr.*, vol. 50, pp. 1745-1762, 1989.
- [8] B. R. Barmish and Z. Shi, "Robust stability of a class of polynomials with coefficients depending multilinearly on perturbations," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 1040-1043, 1990.
- [9] R. R. E. de Gaston and M. G. Safonov, "Exact calculation of the multiloop stability margin," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 156-171, Feb. 1988.
- [10] B. R. Barmish, "New tools for robustness analysis," in *Proc. 27th Conf. Decis. Contr.*, Austin, TX, vol. 1, 1988, pp. 1-6.
- [11] M. P. Polis, A. W. Olbrot, and M. Fu, "An overview of recent results on the parametric approach to robust stability," in *Proc. 28th IEEE Conf. Decis. Contr.*, Tampa, FL, 1989.
- [12] B. R. Barmish, "A generalization of Kharitonov's four polynomial concept for robust stability problems with linear dependent coefficient perturbations," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 157-165, 1989.
- [13] J. J. Anagnost, C. A. Desoer, and R. J. Minnichelli, "Kharitonov's theorem and a graphical stability test for linear time-invariant systems," in *Robustness in Identification and Control*, M. Milanese, R. Tempo, and A. Vicino, Eds. New York: Plenum, 1989.
- [14] S. Dasgupta, P. J. Parker, B. D. O. Anderson, F. J. Kraus, and M. Mansour, "Frequency domain conditions for the robust stability of linear and nonlinear dynamic systems," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 389-397, 1991.
- [15] M. Mansour, F. Kraus, and B. D. O. Anderson, "Strong Kharitonov theorem for discrete systems," in *Proc. 27th Conf. Decision Contr.*, Austin, TX, vol. 1, 1988, pp. 106-111.
- [16] J. D. Cobb and C. L. DeMarco, "The minimal dimension of stable faces required to guarantee stability of a matrix polytope," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 990-992, 1989.
- [17] M. Fu, "Computing the frequency response of a transfer function with parametric perturbations," *Syst. Contr. Lett.*, vol. 15, pp. 45-52, 1990.
- [18] ———, "Polytopes of polynomials with zeros in a prescribed region: New criteria and algorithms," in *Robustness in Identification and Control*, M. Milanese, R. Tempo, and A. Vicino, Eds. New York: Plenum, 1989.
- [19] M. Fu, A. W. Olbrot, and M. P. Polis, "Robust stability for time-delay systems: The edge theorem and graphical tests," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 813-821, 1989.
- [20] G. Chockalingam and S. Dasgupta, "Minimality, stabilizability and strong stabilizability of uncertain plants," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 1651-1661, Nov. 1993.
- [21] A. Rantzer, "A finite zero exclusion principle," in *Proc. MTNS, Amsterdam, Netherlands*, 1989, pp. 239-245.
- [22] I. Horowitz, "Quantitative feedback theory," *Proc. IEE*, part D, vol. 129, pp. 215-226, 1982.
- [23] J. E. Ackermann, "Does it suffice to check a subset of multilinear parameters in robustness analysis?," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 487-488, 1992.
- [24] H. Chapellat, L. H. Keel, and S. P. Bhattacharyya, "Stability margins for multivariable interval control systems," in *Proc. 30th Conf. Decision Contr.*, Brighton, England, 1991, pp. 894-899.
- [25] Y. C. Soh, R. J. Evans, I. R. Petersen, and R. J. Betz, "Robust pole assignment," *Automatica*, vol. 23, no. 5, pp. 601-610, 1987.
- [26] B. T. Polyak, "Robustness analysis of multilinear perturbations," in *Robustness of Dynamic Systems with Parameter Uncertainties*, M. Man-

Sour, S. Balemi, and W. Tuol, Eds. Monte Verita, Birkhauser, Verlag Basel, 1992, pp. 93-104.

B. D. O. Anderson, F. J. Kraus, M. Mansour, and S. Dasgupta, "Easily testable sufficient conditions for the robust stability of systems with multilinear parameter dependence," *Automatica* to be published, 1995.
F. Zeheb, "Necessary and sufficient condition for robust stability of a continuous system—The continuous dependency case illustrated via multilinear dependency," *IEEE Trans. Circuits Syst.* vol. 37, pp. 47-53, Jan. 1990.



Minyue Fu (S'84-M'87) was born in Zhejiang, China, in 1958. He received the B.S. degree in electrical engineering from the China University of Science and Technology, Hefei, China, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from the University of Wisconsin-Madison in 1983 and 1987, respectively.

From 1983-1987 he held a teaching assistantship and a research assistantship at the University of Wisconsin-Madison. In 1987 he was a Computer Engineering Consultant at the Nicolet Instruments

Inc., WI. From 1987-1989 he served as an Assistant Professor in the Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI, where he received an Outstanding Teaching Award. In the summer of 1989 he was a Maitre de Conférences Invited at the Université Catholique de Louvain. Since 1989 he has been with the Department of Electrical and Computer Engineering, University of Newcastle, Australia, where he holds a Senior Lectureship. His current research interests include robust control, dynamical systems, stability, signal processing, and computer engineering.

Dr. Fu was awarded the Maro Guo Scholarship in 1983 for his undergraduate study in China. He is currently an Associate Editor for *TRANSACTIONS ON AUTOMATIC CONTROL*.



Soura Dasgupta (S'81-M'87-SM'90) was born in Calcutta, India. He received a B.S. degree in electrical engineering from the University of Queensland, Australia, in 1980 and a Ph.D. degree in systems engineering from the National University in 1985.

In 1981 he was a Junior Research Fellow at the Department of Electronics and Communications Science, the Indian Statistical Institute, Calcutta. He has had visiting appointments at the University of Notre Dame, University of Iowa, Université Catholique

de Louvain-la-Neuve, Belgium, and the Australian National University. He is currently a Professor with the Department of Electrical and Computer Engineering at the University of Iowa, Iowa City. His current research interests include controls, signal processing, and neural networks.

Dr. Dasgupta served as an Associate Editor for *TRANSACTIONS ON AUTOMATIC CONTROL* from 1988-1991. He is Presidential Faculty Fellow and a corecipient of the Gullimen Cauer Award for the best paper published in the *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS* in the calendar years of 1990 and 1991.



Vincent Blondel was born in Antwerp, Belgium, in 1965. He received the M.Sc. degree in engineering and the Ph.D. degree in applied mathematics from the Catholic University of Louvain in 1988 and 1992, respectively, and the M.Sc. degree in pure mathematics from Imperial College, London, in 1990.

Since 1992, Dr. Blondel has held research positions at the University of Oxford and at the Royal Institute of Technology, Stockholm, where he was the 1993-1994 Göran Gustafsson Research Fellow.

He is currently with INRIA Rocquencourt (the French National Research Institute in Computer Science and Applied Mathematics) near Paris. His current research interests include robust control, linear systems, analytic function theory, and computational complexity of control problems.

Discrete-Time Observers for Singularly Perturbed Continuous-Time Systems

Kenneth R. Shouse, *Member, IEEE*, and David G. Taylor, *Member, IEEE*

Abstract—The use of discrete-time observers for nonlinear singularly perturbed continuous-time systems is explored. The observer design approach is based on inversion of state-to-measurement maps. The two-time-scale nature of the system is exploited by constructing separate reduced-order observers for the approximate slow and fast subsystems, using multirate measurements and computation. The reduced-order observers are compared to a full-order observer designed for the same system. The comparison shows that although the reduced-order observers exhibit some approximation error, they also result in reduced computational requirements if the inversion is performed on-line, reduced memory requirements if the inversion is implemented by look-up table and reduced stiffness. The trade-off between accuracy and implementation requirements always favors the reduced-order approach for systems with significant separation of time scales. Numerical examples are provided, including a case-study of an observer-based control system for permanent-magnet synchronous motors.

I. INTRODUCTION

IN this paper, the problem of constructing state observers for nonlinear continuous-time singularly perturbed systems is investigated. One observer design concept that may be applied to such systems is based on inversion of the so-called state-to-measurement map (a set of algebraic equations) [4], [5]. Since this design concept is widely applicable, requiring only that standard observability assumptions be satisfied by the discrete-time representation of the system, this paper will adopt a map-inversion approach and will adhere to the basic principles outlined in [4], [5].

Unfortunately, when this full-order observer is applied to singularly perturbed systems, several difficulties are encountered. First, the discretization of the continuous-time system must be carried out at a sufficiently small sampling interval to accurately model the fast states, even though the states of primary interest may be the slow ones. This necessitates the use of faster sampling devices and faster processors than would otherwise be needed. Second, because of this separation in time scales, the equations which must be inverted to construct the observer are stiff. This means that accurate numerical solution is more difficult than if the system evolved in a single time scale.

Manuscript received June 12, 1992; revised May 20, 1993. Recommended by Past Associate Editor, A. Arapostathis. This work was supported in part by National Science Foundation Grants ECS-8909329, ECS-9007778, and ECS-9158037 and by Clifton Precision South, a division of Litton Systems, Inc.

K. R. Shouse is with the Engine and Vehicle Research Division, Southwest Research Institute, San Antonio, TX 78228-0510 USA.

D. G. Taylor is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250 USA.

IEEE Log Number 9407564.

The two-time-scale nature of the system can be exploited by decomposing the problem into one of designing separate reduced-order observers for the approximate slow and fast subsystems. This decomposition is quite straightforward for systems without inputs, but can also be applied when inputs are present with some technical modifications. The outputs of the two reduced-order observers can be used to estimate the state of the original system. Since the slow and fast observers are separate algorithms, they may be implemented using multirate measurements and computation. Consequently, the computational and/or memory requirements for state-to-measurement map inversion are significantly reduced from those of the full-order observer, and the numerical stiffness is relieved.

All the benefits of the reduced-order observers are achieved through a trade-off with accuracy. The reduced-order observers are only approximate, but each source of observation error vanishes as the singular perturbation parameter tends to zero. To demonstrate the practical value of reduced-order observers for feedback control, the paper concludes with a detailed case-study on velocity regulation of permanent-magnet synchronous motors with only stator current measurements.

II. FULL-ORDER ANALYSIS

In this section, the full-order observer for singularly perturbed autonomous systems is analyzed. This analysis is presented to provide a reference for comparison with the reduced-order observers that follow.

A The Full-Order Observer

We follow precisely the procedure recommended in [4], [5]. Consider the continuous-time nonlinear system

$$\dot{x} = f_1(x, z) \quad (2.1)$$

$$\dot{z} = f_2(x, z) \quad (2.2)$$

$$y = h(x, z) \quad (2.3)$$

where $x \in \mathcal{R}^N$, $z \in \mathcal{R}^M$, $y \in \mathcal{R}^Q$, $\epsilon > 0$ is a small parameter and the dot denotes differentiation with respect to time t . It is assumed that f_1 , f_2 and h are analytic, that unique solutions $x(t)$ and $z(t)$ exist, and that y is the only measured quantity. For notational convenience the representation

$$\dot{w} = f(w) := \begin{bmatrix} f_1(w_1, w_2) \\ \frac{1}{\epsilon} f_2(w_1, w_2) \end{bmatrix} \quad (2.4)$$

$$y = h(w) := h(w_1, w_2) \quad (2.5)$$

is used, where $w = [w_1' w_2']'$ with $w_1 := x$ and $w_2 := z$.

Discretizing this system at the instants $t = \ell\epsilon T$ where T is suitably chosen on the basis of the x dynamics, one obtains

$$w[\ell + 1] = F(w[\ell]) \quad (2.6)$$

$$y[\ell] = h(w[\ell]) \quad (2.7)$$

here $w[\ell] := w(\ell\epsilon T)$, $y[\ell] := y(\ell\epsilon T)$

$$F(w[\ell]) := \sum_{i=0}^{\infty} \frac{(\epsilon T)^i}{i!} [L_f^i w]_{w=u[\ell]} \quad (2.8)$$

and where the notation $L_f^i w$ denotes the Lie derivative operator L_f^i applied to each component of w [10].

Associated with the system (2.6)–(2.7) is the full-order state-to-measurement map

$$H(w) := \begin{bmatrix} h(w) \\ h \circ F(w) \\ \vdots \\ h \circ F^{N+M-1}(w) \end{bmatrix} \quad (2.9)$$

where F^i denotes F composed with itself i times. This state-to-measurement map satisfies

$$Y[\ell] = H(w[\ell - N - M + 1]) \quad (2.10)$$

where $Y[\ell]$ is the block of measurements

$$Y[\ell] := \begin{bmatrix} y[\ell - N - M + 1] \\ \vdots \\ y[\ell] \end{bmatrix} \quad (2.11)$$

Assumption 2.1 For some set $D_u \subseteq \mathcal{R}^{N+M}$, there exists a function $H^{-1}: \mathcal{R}^{(N+M)Q} \rightarrow \mathcal{R}^{N+M}$ such that $w \in D_u$ and $Y = H(w)$ imply $w = H^{-1}(Y)$.

Under Assumption 2.1, the full-order observer for the system (2.6)–(2.7) [and hence for the system (2.4)–(2.5)] is given by

$$\hat{w}[\ell] := F^{N+M-1} \circ H^{-1}(Y[\ell]) \quad (2.12)$$

This nonlinear observer is deadbeat on D_u , i.e., $w[\ell] = \hat{w}[\ell]$ for all $\ell \geq N + M - 1$ if $w[\ell - N - M + 1] \in D_u$.

To gain insight into the observer design given above and to aid in the analysis to follow, the specialization of this observer to the linear autonomous case is useful. Consider the system (2.4)–(2.5) defined by $f(w) = Aw$ and $h(w) = C'w$, where

$$A := \begin{bmatrix} A_{11} & A_{12} \\ \frac{1}{\epsilon}A_{21} & \frac{1}{\epsilon}A_{22} \end{bmatrix} \quad (2.13)$$

$$C' := [C'_1 \quad C'_2]. \quad (2.14)$$

Discretizing at the instants $t = \ell\epsilon T$ yields the model (2.6)–(2.7) with $F(w[\ell]) = \Phi w[\ell]$ and $h(w[\ell]) = C'w[\ell]$ where $\Phi := \exp(A\epsilon T)$. The specialization of Assumption 2.1 simply requires that the observability matrix

$$\mathcal{O} := \begin{bmatrix} C' \\ C'\Phi \\ \vdots \\ C'\Phi^{N+M-1} \end{bmatrix} \quad (2.15)$$

have full column rank. In this case, the full order observer

$$\hat{w}[\ell] := \Phi^{N+M-1}(\mathcal{O}'\mathcal{O})^{-1}\mathcal{O}'Y[\ell]$$

which has the deadbeat property $\hat{w}[\ell] = w[\ell]$ for $\ell \geq N + M - 1$.

B Full-Order Observer Stiffness

The full-order observer for a singularly perturbed system is inherently stiff. To see this, consider the linear system (2.13)–(2.14). To this system, apply the diagonalizing transformation [8]

$$\begin{bmatrix} \xi \\ \eta \end{bmatrix} := \begin{bmatrix} I_N - \epsilon L_2 L_1 & -\epsilon L_2 \\ L_1 & I_M \end{bmatrix} \begin{bmatrix} r \\ z \end{bmatrix} \quad (2.17)$$

where L_1 and L_2 solve

$$A_{21} - A_{22}L_1 + \epsilon L_1 A_{11} - \epsilon L_1 A_{12}L_1 = 0 \quad (2.18)$$

$$\epsilon(A_{11} - A_{12}L_1)L_2 - L_2(A_{22} + \epsilon L_1 A_{12}) + A_{12} = 0 \quad (2.19)$$

Note that L_1 and L_2 are guaranteed to exist provided that A_{22} is nonsingular (i.e., that the system is in “standard” form), and that ϵ is sufficiently small [8]. The result is

$$\begin{bmatrix} \dot{\xi} \\ \dot{\eta} \end{bmatrix} = \begin{bmatrix} A_\xi & 0 \\ 0 & A_\eta \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix} \quad (2.20)$$

$$y = [C'_\xi \quad C'_\eta] \begin{bmatrix} \xi \\ \eta \end{bmatrix} \quad (2.21)$$

where $A_\xi := A_{11} - A_{12}L_1$, $A_\eta := A_{22} + \epsilon L_1 A_{12}$, $C'_\xi := C'_1 - C'_2 L_1$ and $C'_\eta := C'_2 - \epsilon(C'_1 L_2 - C'_2 L_1 L_2)$. Discretizing at the instants $t = \ell\epsilon T$ gives

$$\begin{bmatrix} \xi[\ell + 1] \\ \eta[\ell + 1] \end{bmatrix} = \begin{bmatrix} \Phi_\xi & 0 \\ 0 & \Phi_\eta \end{bmatrix} \begin{bmatrix} \xi[\ell] \\ \eta[\ell] \end{bmatrix} \quad (2.22)$$

$$y[\ell] = [C'_\xi \quad C'_\eta] \begin{bmatrix} \xi[\ell] \\ \eta[\ell] \end{bmatrix} \quad (2.23)$$

where $\Phi_\xi := \exp(A_\xi \epsilon T) = I_N + O(\epsilon)$ and $\Phi_\eta := \exp(A_\eta \epsilon T)$.

In these new coordinates, the observability matrix (2.15) is given by

$$\mathcal{O}_d := \begin{bmatrix} C'_\xi & C'_\eta \\ C'_\xi + O(\epsilon) & C'_\eta \Phi_\eta \\ \vdots & \vdots \\ C'_\xi + O(\epsilon) & C'_\eta \Phi_\eta^{N+M-1} \end{bmatrix}. \quad (2.24)$$

Since the first N columns of \mathcal{O}_d are nearly the same for small ϵ , it is very close to losing rank. The observability matrix (2.15) in original coordinates is

$$\mathcal{O} = \mathcal{O}_d \begin{bmatrix} I_N - \epsilon L_2 L_1 & -\epsilon L_2 \\ L_1 & I_M \end{bmatrix}. \quad (2.25)$$

Since the transformation matrix in (2.25) does not lose rank as $\epsilon \rightarrow 0$, \mathcal{O} is also ill-conditioned for small ϵ .

In the nonlinear case, ill-conditioning will be evident whenever $\partial H / \partial w$ is close to losing rank. Thus Jacobian clearly is the observability matrix of the classical linearization of the nonlinear discrete-time system (2.6)–(2.7). Since classical linearization and discretization are commutative [3], however,

this Jacobian is also the observability matrix of the discretization of some linearization of the system (2.1)–(2.3). Since any such observability matrix is inherently ill-conditioned, it follows that the nonlinear observer (2.12) is inherently stiff.

The stiffness of the full-order observer is a result of the inordinately fast sampling required. This fast sampling causes an ϵ dependence in $H(w)$, which leads to loss of rank as $\epsilon \rightarrow 0$. Since reduced-order observers may be operated in their natural time scales (in a multirate fashion), such stiffness is not an inherent feature of the approach considered in the following sections.

III. REDUCED-ORDER DESIGN

The problem of observer design for autonomous nonlinear singularly perturbed systems is now re-examined. The proposed reduced-order techniques can be applied to most singularly perturbed systems for which a full-order observer exists.

A. Time-Scale Decomposition

Appropriate design models are needed for the construction of reduced-order observers. These models will be determined by first obtaining continuous-time reduced-order models and by then computing the corresponding discrete-time reduced-order models. This is in contrast to [1], where full-order discretization is followed by discrete-time order-reduction. By performing the order reduction first, no tools from the (less developed) discrete-time singular perturbation theory are needed. By discretizing only models of reduced order, the resulting state-to-measurement maps (and their inverses) will be simpler.

Consider systems of the form (2.1)–(2.3), namely

$$\dot{x} = f_1(x, z), x(0) = x_0 \quad (3.1)$$

$$\epsilon \dot{z} = f_2(x, z), z(0) = z_0 \quad (3.2)$$

$$y = h(x, z) \quad (3.3)$$

where all notation and assumptions match those of (2.1)–(2.3). It is implicitly assumed that f_1 and f_2 have $O(1)$ magnitudes, so that the rate-of-change of z is about $1/\epsilon$ times larger than the rate-of-change of x . It will be convenient in the sequel to make use of the set notation $\mathcal{B}_r(\delta) := \{\chi \in \mathcal{R}^J : \|\chi\|_\infty < \delta\}$.

Assumption 3.1: For a fixed $t > 0$, there exist fixed positive scalars r_x and r_z such that $(x(t), z(t)) \in \mathcal{B}_N(r_x) \times \mathcal{B}_M(r_z)$ for all $t \in [0, t]$ and such that $0 = f_2(x, z)$ has the unique real solution $z = \phi(x)$ on $(x, z) \in \mathcal{B}_N(r_x) \times \mathcal{B}_M(r_z)$.

Following [8], the slow subsystem is defined by

$$\frac{dx_s}{dt} = f_1(x_s, \phi(x_s)) =: f_s(x_s), x_s(0) = x_0 \quad (3.4)$$

$$y_s = h(x_s, \phi(x_s)) =: h_s(x_s) \quad (3.5)$$

and the fast subsystem is defined by

$$\frac{dz_f}{d\tau} = f_2(x_0, \phi(x_0) + z_f) =: f_f(z_f), z_f(0) = z_0 - \phi(x_0) \quad (3.6)$$

$$y_f = h(x_0, \phi(x_0) + z_f) - h(x_0, \phi(x_0)) =: h_f(z_f) \quad (3.7)$$

where $\tau := t/\epsilon$. It is assumed that a unique solution $x_s(t)$ of the system (3.4)–(3.5) exists for all $t \in [0, t]$, and a unique solution $z_f(\tau)$ of the system (3.6)–(3.7) exists for all $\tau \in [0, t/\epsilon]$.

It is significant that $f_f(z_f)$ and $h_f(z_f)$ depend implicitly on the initial condition x_0 . This feature is inherent in the time-scale decomposition. It should also be emphasized that y_s and y_f are not signals available for observer design purposes.

Assumption 3.2: The fast states are asymptotically stable, i.e. $\text{Re}\left\{\lambda\left(\frac{\partial f_f}{\partial z}\right)\right\} \leq -\kappa < 0$ evaluated along $x = x_s$ and $z = \phi(x_s)$, for some fixed constant $\kappa > 0$. Furthermore, $x(0) = x_0$ and $z(0) = z_0$ are such that $z_f(0)$ belongs to the domain of attraction of the equilibrium $z_f = 0$ of the fast subsystem.

The following well-known theorem provides a method of approximating $x(t)$, $z(t)$ and $y(t)$ by solving the reduced-order subsystems (3.4)–(3.7) [7], [8].

Theorem 3.1 [Tikhonov]: If the system (3.1)–(3.3) satisfies Assumptions 3.1 and 3.2, then there exists an $\epsilon^* > 0$ such that for any $\epsilon \in (0, \epsilon^*)$

$$x(t) = x_s(t) + O(\epsilon) \quad (3.8)$$

$$z(t) = \phi(x_s(t)) + z_f(\tau) + O(\epsilon) \quad (3.9)$$

$$y(t) = y_s(t) + y_f(\tau) + O(\epsilon) \quad (3.10)$$

for all $t \in [0, t]$. Furthermore, for any $t^* \in (0, t]$ there exists an $\epsilon^{**} \in (0, \epsilon^*)$ such that for any $\epsilon \in (0, \epsilon^{**})$

$$z(t) = \phi(x_s(t)) + O(\epsilon) \quad (3.11)$$

for all $t \in [t^*, t]$

B. Synthesis of Reduced-Order Observers

Since the reduced-order models (3.4)–(3.5) and (3.6)–(3.7) are decoupled, they may be sampled at their natural rates. The slow subsystem is sampled at the instants $t = nT$ and the fast subsystem at the instants $\tau = mT/(k\epsilon)$, where $T \leq t/N$ is chosen on the basis of the x dynamics and where $k \geq M$ is an integer such that $T/k = O(\epsilon)$. Note that the fast sampling rate is precisely k times the slow sampling rate, and also that the separation in time scales is maintained in the discrete n and m indexes. Throughout this section, time-varying quantities will be written as functions of t (or n), or τ (or m), in a manner consistent with the notation of Theorem 3.1.

The discrete-time model of the slow subsystem is

$$x_s[n+1] = F_s(x_s[n]) \quad (3.12)$$

$$y_s[n] = h_s(x_s[n]) \quad (3.13)$$

where $x_s[n] := x_s(nT)$, $y_s[n] := y_s(nT)$, and

$$F_s(x_s[n]) := \sum_{i=0}^{T/\epsilon} \frac{T^i}{i!} [L_{f_s}^i x_s]_{x_s=x_s[n]}. \quad (3.14)$$

The discrete-time model of the fast subsystem is

$$z_f[m+1] = F_f(z_f[m]) \quad (3.15)$$

$$y_f[m] = h_f(z_f[m]) \quad (3.16)$$

where $z_f[m] := z_f(mT/(k\epsilon))$, $y_f[m] := y_f(mT/(k\epsilon))$, and

$$F_f(z_f[m]) := \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{T}{k\epsilon} \right)^i \left[L_{f_f}^i z_f \right]_{z_f=z_f[m]} \quad (3.17)$$

It may seem that constructing observers for the reduced-order models (3.12)–(3.13) and (3.15)–(3.16) is simply a matter of repeating the full-order procedure for each of the subsystems. This is not possible, however, because only $y(t)$ is measured— $y_s[n]$ and $y_f[m]$ are not known. Intuitively, fast variations in the measured output may be attributed to the fast states. Consequently, any difference between $y(mT/k)$ and $y((m+1)T/k)$ will be approximately equal to the difference between $y_f[m]$ and $y_f[m+1]$. For this reason, the construction of the fast observer begins with the fast output function in incremental form

$$\Delta h_f(z_f) := h_f \circ F_f(z_f) - h_f(z_f). \quad (3.18)$$

Using (3.18), the incremental fast state-to-measurement map is defined as

$$\Delta H_f(z_f) := \begin{bmatrix} \Delta h_f(z_f) \\ \Delta h_f \circ F_f(z_f) \\ \vdots \\ \Delta h_f \circ F_f^{M-1}(z_f) \end{bmatrix} \quad (3.19)$$

which is implicitly parameterized by x_0 . Given the block of incremental fast subsystem outputs

$$\Delta Y_f[m] := \begin{bmatrix} y_f[m-M+1] - y_f[m-M] \\ \vdots \\ y_f[m] - y_f[m-1] \end{bmatrix} \quad (3.20)$$

this map satisfies the identity

$$\Delta Y_f[m] = \Delta H_f(z_f[m-M]). \quad (3.21)$$

Assumption 3.3: For some fixed positive scalars r_f and δ_f , there exists a function $\Delta H_f^{-1} : \mathcal{R}^{QM} \rightarrow \mathcal{R}^M$, parameterized by x_0 , such that:

- 1) $z_f \in \mathcal{B}_M(r_f)$ and $\Delta Y_f = \Delta H_f(z_f)$ imply $z_f = \Delta H_f^{-1}(\Delta Y_f)$.
- 2) $\Delta H_f^{-1}(\cdot)$ depends continuously on the parameter x_0 and is continuous on the set $\mathcal{B}_{QM}(r + \delta_f)$ where $r > 0$ is such that $\Delta H_f(\mathcal{B}_M(r_f)) \subseteq \mathcal{B}_{QM}(r)$.

Assumption 3.3–2 is required because of the approximation error introduced by order reduction. This requirement is not restrictive, since a similar assumption would be required in [4], [5] to handle measurement noise and/or parameter uncertainty. Invoking Assumption 3.3, whenever $z_f[m-M] \in \mathcal{B}_M(r_f)$ then

$$z_f[m] = F_f^M \circ \Delta H_f^{-1}(\Delta Y_f[m]) \quad (3.22)$$

which is computable for $m \geq M$.

Turning to the slow observer problem, consider the sum of blocks of slow and fast subsystem outputs

$$Y_{sf}[n] := \begin{bmatrix} y_s[n-N+1] \\ \vdots \\ y_s[n] \end{bmatrix} + \begin{bmatrix} y_f[k(n-N+1)] \\ \vdots \\ y_f[kn] \end{bmatrix}. \quad (3.23)$$

This quantity motivates the slow-plus-fast slow measurement map¹ definition

$$H_{sf}(x_s, Z_f[n]) := \begin{bmatrix} h_s(x_s) \\ h_s \circ F_s(x_s) \\ \vdots \\ h_s \circ F_s^{N-1}(x_s) \\ h_f(z_f[k(n-N+1)]) \\ h_f(z_f[k(n-N+2)]) \\ \vdots \\ h_f(z_f[kn]) \end{bmatrix} \quad (3.24)$$

which is parameterized implicitly by x_0 and explicitly by

$$Z_f[n] := \begin{bmatrix} z_f[k(n-N+1)] \\ \vdots \\ z_f[kn] \end{bmatrix}. \quad (3.25)$$

This map satisfies the identity

$$Y_{sf}[n] = H_{sf}(x_s[n-N+1], Z_f[n]). \quad (3.26)$$

Assumption 3.4: For the fixed positive scalar r_f and some fixed positive scalars r_s and δ_s , there exists a function $H_{sf}^{-1} : \mathcal{R}^{QN} \times \mathcal{R}^{MN} \rightarrow \mathcal{R}^N$ such that:

- 1) $x_s \in \mathcal{B}_N(r_s)$ and $Y_{sf} = H_{sf}(x_s, Z_f)$ imply $x_s = H_{sf}^{-1}(Y_{sf}, Z_f)$ for all $Z_f \in \mathcal{B}_{MN}(r_f)$.
- 2) $H_{sf}^{-1}(\cdot, \cdot)$ is continuous on the set $\mathcal{B}_{QN+MN}(r + \delta_s)$ where $r > 0$ is such that $H_{sf}(\mathcal{B}_N(r_s), \mathcal{B}_{MN}(r_f)) \times \mathcal{B}_{MN}(r_f) \subseteq \mathcal{B}_{QN+MN}(r)$.

Again, Assumption 3.4-2 is required to account for order reduction approximation error. Invoking Assumption 3.4, whenever $x_s[n-N+1] \in \mathcal{B}_N(r_s)$ and $Z_f[n] \in \mathcal{B}_{MN}(r_f)$, then

$$x_s[n] = F_s^{N-1} \circ H_{sf}^{-1}(Y_{sf}[n], Z_f[n]) \quad (3.27)$$

which is computable for $n \geq N-1$.

Of course, neither (3.22) nor (3.27) is computable using known signals. The fictitious signals ΔY_f and Y_{sf} may be approximated by

$$\Delta \hat{Y}_f[m] := \begin{bmatrix} y((m-M+1)T/k) - y((m-M)T/k) \\ \vdots \\ y(mT/k) - y((m-1)T/k) \end{bmatrix} \quad (3.28)$$

$$\hat{Y}_{sf}[n] := \begin{bmatrix} y((n-N+1)T) \\ \vdots \\ y(nT) \end{bmatrix} \quad (3.29)$$

and the unknown $Z_f[n]$ will be obtained from the fast observer defined below, according to

$$\hat{Z}_f[n] := \begin{bmatrix} \hat{z}_f[k(n-N+1)] \\ \vdots \\ \hat{z}_f[kn] \end{bmatrix}. \quad (3.30)$$

¹ Although in this section it is possible to formulate an implementable slow observer using a "slow" state-to-measurement map, this definition has been chosen to be consistent with a similar definition required in the next section.

TABLE I
 IMPLEMENTATION COSTS OF FULL- AND REDUCED-ORDER OBSERVERS

Observer	Numerical (flops)	Closed-Form (flops)	Look-Up Table (bytes)
Full-Order	$\frac{O((N+M)^3)}{T}$	$\frac{O((N+M)^2)}{T}$	$2^{b(N+M)+2}$
Theorem 3.2	$\frac{O(M^3)}{T} + \frac{O(N^3)}{T}$	$\frac{O(M^2)}{T} + \frac{O(N^2)}{T}$	$2^{bM+2} + 2^{bN+2}$
Corollary 3.1	$\frac{O(N^3)}{T}$	$\frac{O(N^2)}{T}$	2^{bN+2}

Note also that F_f , ΔH_f^{-1} and H_{sf}^{-1} all depend implicitly on x_0 . Hence, the following assumption is required.

Assumption 3.5: At least one of the following conditions is satisfied:

- 1) $\hat{x}_0 = x_0 + O(\epsilon)$ is known.
- 2) $f_2(x, z) = f_{2x}(x) + A_2 z$ and $h(x, z) = h_x(x) + C z$.

If Assumption 3.5-2 is true, then F_f , ΔH_f^{-1} , and H_{sf}^{-1} are independent of x_0 , since $f_f(z_f) = A_2 z_f$ and $h_f(z_f) = C z_f$. If Assumption 3.5-1 is true, then the unknown functions will be estimated by

$$\hat{F}_f(z_f) := F_f(z_f)|_{x_0=\hat{x}_0} \quad (3.31)$$

$$\Delta \hat{H}_f^{-1}(z_f) := \Delta H_f^{-1}(z_f)|_{x_0=\hat{x}_0} \quad (3.32)$$

$$\hat{H}_{sf}^{-1}(x_s, Z_f) := H_{sf}^{-1}(x_s, Z_f)|_{x_0=\hat{x}_0} \quad (3.33)$$

The implementable reduced-order observers are defined in the following theorem, and they are interconnected in the sense that $\hat{z}_f[m]$ drives $\hat{x}_s[n]$.

Theorem 3.2: Suppose $(x_s[n], z_f[m]) \in \mathcal{B}_N(r_s) \times \mathcal{B}_M(r_f)$ for $n = 0, \dots, \bar{n}$ and $m = 0, \dots, k\bar{n}$, where \bar{n} is the greatest integer such that $\bar{n}T \leq t$. If the system (3.1)–(3.3) satisfies Assumptions 3.1–3.5, then there exists an $\epsilon^* > 0$ such that for any $\epsilon \in (0, \epsilon^*)$

$$\hat{z}_f[m] := \hat{F}_f^M \circ \Delta \hat{H}_f^{-1}(\Delta \hat{Y}_f[m]) \quad (3.34)$$

$$= z_f(mT/(k\epsilon)) + O(\epsilon) \quad (3.35)$$

$$\hat{x}_s[n] := F_s^{N-1} \circ H_{sf}^{-1}(\hat{Y}_{sf}[n], \hat{Z}_f[n]) \quad (3.36)$$

$$= x_s(nT) + O(\epsilon) \quad (3.37)$$

for $n = N, \dots, n$ and $m = M, \dots, kn$.

Proof: Assumptions 3.1 and 3.2 imply that Theorem 3.1 holds for $n = 0, \dots, n$ (and $m = 0, \dots, k\bar{n}$). In light of (3.4) and the fact $T/k = O(\epsilon)$, it is evident that $r_s((m+1)T/k) = x_s(mT/k) + O(\epsilon)$. Since h_s is analytic, it is continuous and therefore $y_s((m+1)T/k) = y_s(mT/k) + O(\epsilon)$, which, along with (3.10), implies that

$$\Delta Y_f[m] = \Delta Y_f[m] + O(\epsilon). \quad (3.38)$$

This, together with Assumptions 3.3 and 3.5, the continuity of F_f and the hypothesis that $z_f[m] \in \mathcal{B}_M(r_f)$, implies (3.35). From (3.10), it follows that

$$\hat{Y}_{sf}[n] = Y_{sf}[n] + O(\epsilon). \quad (3.39)$$

Using (3.35), it follows that

$$\hat{Z}_f[n] = Z_f[n] + O(\epsilon). \quad (3.40)$$

Along with Assumption 3.4, the hypothesis that $x_s[n] \in \mathcal{B}_N(r_s)$ and the continuity of F_s , (3.39)–(3.40) imply (3.37). \square

Since the observers (3.34) and (3.36) are interconnected, it is helpful to describe their implementation further. The required sequence of calculations is as follows:

- 1) Using \hat{x}_0 if necessary, construct (3.31), (3.32) and (3.33).
- 2) Set $m = n = 0$.
- 3) While $n \leq \bar{n}$
 - a) Measure $y(mT/k)$.
 - b) If $m \geq M$, compute $\hat{z}_f[m]$ from (3.34) using $y((m-M)T/k), \dots, y(mT/k)$.
 - c) If $m = k(n+1)$ then
 - Set $n = n + 1$.
 - If $n \geq N$, compute $\hat{x}_s[n]$ from (3.36) using $y((n-N+1)T), \dots, y(nT)$ and $\hat{z}_f[k(n-N+1)], \dots, \hat{z}_f[kn]$.
 - d) Set $m = m + 1$.
- 4) Stop.

If the accuracy of the observer during the boundary layer is of no concern, then an estimate of $z_f[m]$ is not needed, and it is possible to obtain a simplified slow observer.

Corollary 3.1. Suppose that $x_s[n] \in \mathcal{B}_N(r_s)$ for $n = 0, \dots, n$. If the system (3.1)–(3.3) satisfies Assumptions 3.1, 3.2, and 3.4, then for any integer $n^* \in [N, \bar{n}]$ there exists an $\epsilon^{**} > 0$ such that for any $\epsilon \in (0, \epsilon^{**})$

$$\hat{x}_s[n] := F_s^{N-1} \circ H_{sf}^{-1}(\hat{Y}_{sf}[n], 0) \quad (3.41)$$

$$= x_s(nT) + O(\epsilon) \quad (3.42)$$

for all $n = n^*, \dots, n$.

Proof: Equation (3.11) implies that, under the hypothesis of the corollary, $z_f[m] = O(\epsilon)$ for $m \in [kt^*/T, kn]$. Taking n^* as the smallest integer such that $n^* \geq N - 1 + t^*/T$ leads to $Z_f[n] = O(\epsilon)$ for $n = n^*, \dots, n$. This, together with the fact that $\hat{H}_{sf}(\cdot, 0) \equiv H_{sf}(\cdot, 0)$ (which is evident from (3.24)) and (3.36), proves the result. \square

Corollary 3.1 reinforces the intuitive notion that if the parasitic states are fast enough, then their effects can be completely neglected. Note also that the observer (3.41) does not require either the invertibility of the incremental fast state-to-measurement map, knowledge of x_0 , or special structure of f_2 and h .

C. Reduced-Order Observer Advantages

Table I provides a comparison of the implementation requirements for both full-order and reduced-order observers, assuming $k = 1/\epsilon$. It is further assumed that numerical solution of j nonlinear equations requires $O(j^3)$ floating point operations, and that evaluation of a closed form solution to j nonlinear equations requires $O(j^2)$ floating point operations. If the inverse of a j -dimensional map is computed off-line and stored in a single precision floating point look-up table with b bit resolution, then 2^{bj+2} bytes of memory are assumed to be required. As a consequence of the conservative assumptions used, the values shown in Table I underestimate the advantages of the reduced-order approach.

To better illustrate these requirements, consider a fourth order system with two slow and two fast states ($N = M = 2$). The limiting ratio of on-line numerical inversion computational requirements is

$$\lim_{\epsilon \rightarrow 0} \frac{\text{full-order flops}}{\text{reduced-order flops}} = \begin{cases} 8, & \text{Theorem 3.2} \\ \infty, & \text{Corollary 3.1} \end{cases}$$

The 8-bit look-up table memory requirements for the full- and reduced-order implementations for this system are given by

$$\text{memory size} = \begin{cases} 16 \text{ Gbyte,} & \text{full-order} \\ 512 \text{ Kbyte,} & \text{Theorem 3.2} \\ 256 \text{ Kbyte,} & \text{Corollary 3.1.} \end{cases}$$

Clearly, a look-up table full-order observer is not practical for this system, whereas both of the reduced-order schemes require only moderately sized look-up tables.

As for the stiffness, there is no assurance that the reduced-order inversions will be well-conditioned, since poor observability will not be improved by singular perturbation approximations. On the other hand, as previously shown, the full-order inversions are always ill-conditioned for small enough ϵ .

D. Autonomous Example

Consider the linear system

$$\dot{x} = \begin{bmatrix} 30.9265 & -24.9047 \\ 35.3095 & -27.4703 \end{bmatrix} x + \begin{bmatrix} -13.9306 & 23.3253 \\ -16.8008 & 25.4823 \end{bmatrix} z \quad (3.43)$$

$$\epsilon \dot{z} = \begin{bmatrix} 24.5739 & -18.8923 \\ 9.4471 & -6.1202 \end{bmatrix} x + \begin{bmatrix} -12.2427 & 19.0368 \\ -5.9734 & 6.2865 \end{bmatrix} z \quad (3.44)$$

$$y = [1.5630 \quad 1.6125]x + [1.7635 \quad 2.7318]z \quad (3.45)$$

where $x_1(0) = 1$, $x_2(0) = 2$, $z_1(0) = -2$, $z_2(0) = 1$, $T = 0.05$, and k is the greatest integer less than or equal to $1/\epsilon$. This system satisfies the requirements of Assumptions 3.1–3.5. According to Theorem 3.2, when $1/\epsilon$ is an integer, the fast observer is given by

$$\hat{z}_f[m] = \begin{bmatrix} 2.5922 & -3.4629 \\ 1.7732 & -2.0408 \end{bmatrix} \begin{bmatrix} y[m-1] - y[m-2] \\ y[m] - y[m-1] \end{bmatrix} \quad (3.46)$$

and, otherwise, the numerical matrix in this expression will be a perturbation of that shown. The slow observer is given by

$$\hat{x}_s[n] = \begin{bmatrix} -1.7104 & 1.9170 \\ 0.1755 & 0.0236 \end{bmatrix} \begin{bmatrix} y[n-1] - [1.7635 \quad 2.7318]\hat{z}_f[k(n-1)] \\ y[n] - [1.7635 \quad 2.7318]\hat{z}_f[kn] \end{bmatrix} \quad (3.47)$$

Fig. 1 shows the actual and reduced-order estimated responses for the case $\epsilon = 0.04$. The reduced-order observers do an excellent job of reconstructing the slow and fast states for this value of ϵ . The behavior of the observers as ϵ varies from 0.01 to 0.1 is shown in Fig. 2. In Fig. 2(a), the averages (over the time indexes) of the absolute values of the observer errors are shown. For small values of ϵ the estimates are good, but for large enough ϵ , the estimates degrade as expected.

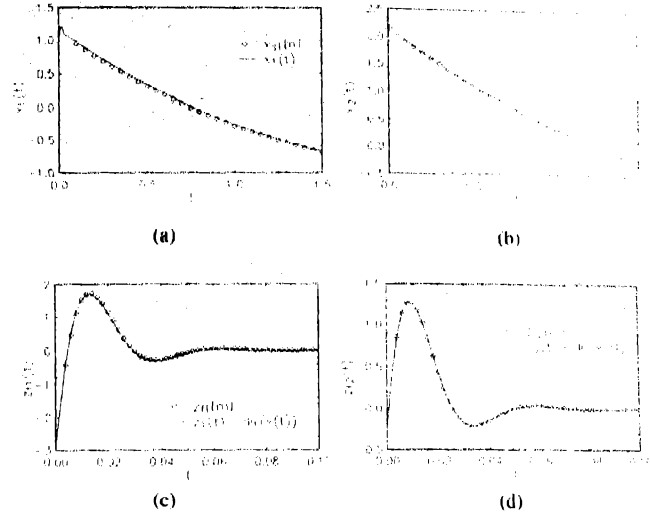


Fig. 1. Actual and estimated state responses for the case $\epsilon = 0.04$. (a) $x_1(t)$ and $\hat{x}_{s1}[n]$, (b) $x_2(t)$ and $\hat{x}_{s2}[n]$, (c) $z_1(t) - \phi_1(x(t))$ and $\hat{z}_{f1}[m]$ and (d) $z_2(t) - \phi_2(x(t))$ and $\hat{z}_{f2}[m]$.

Fig. 2(b) compares the matrix 2-norm condition numbers of the full-order observability matrix, the slow observability matrix, and the incremental fast observability matrix. The full-order observer for this system is ill-conditioned—particularly for small ϵ , whereas the reduced-order observers are not. Fig. 2(c) shows the ratio of the full-order to reduced-order flop count. The reduced-order observers require only about 40% of the calculations required by the full-order observer, assuming that the matrix inversions are computed off-line. Based on the earlier analysis, it is expected that this flops ratio will approach four as $\epsilon \rightarrow 0$.

IV. REDUCED-ORDER DESIGN WITH INPUTS

In this section, the design of reduced-order observers is extended to a class of nonlinear singularly perturbed systems with piecewise-constant inputs. The section concludes with a practical application of the theory: velocity regulation of permanent-magnet synchronous motors, using only stator current measurements.

A. Decomposition with Piecewise-Constant Inputs

Nonlinear singularly perturbed systems exhibiting linearity in the fast states and the control inputs admit reduced-order models that are easy to discretize. Moreover, this class of systems also models the dynamics of some important electromechanical devices. Hence, we consider systems of the form

$$\dot{x} = a_1(x) + A_1(x)z + B_1(x)u, x(0) = x_0 \quad (4.1)$$

$$\epsilon \dot{z} = a_2(x) + A_2(x)z + B_2(x)u, z(0) = z_0 \quad (4.2)$$

$$y = c(x) + C(x)z \quad (4.3)$$

where $x \in \mathcal{R}^N$, $z \in \mathcal{R}^M$, $u \in \mathcal{R}^P$, $y \in \mathcal{R}^Q$, $\epsilon > 0$ is a small parameter, and the dot denotes differentiation with respect to t . It is assumed that $a_1, a_2, A_1, A_2, B_1, B_2, c$ and C are analytic functions, that unique solutions $x(t)$ and $z(t)$ exist, and that y is the only measurement available. Furthermore,

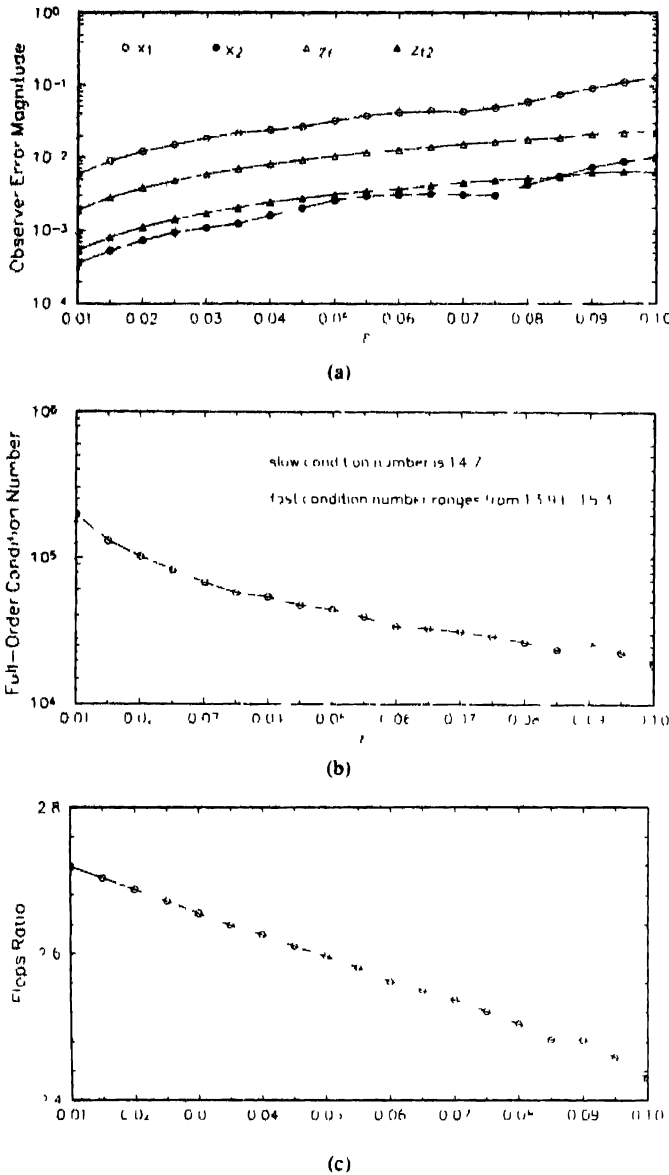


Fig. 2 Observer performance as a function of ϵ (a) Average absolute observer error, (b) Observability matrix condition numbers, and (c) Full-order to reduced-order flops ratio

$u(t)$ is assumed to be piecewise-constant

$$u(t) = u(nT) =: u[n], \forall t \in [nT, (n+1)T) \quad (4.4)$$

where T is consistent with the speed of the x dynamics. The need to accommodate piecewise-constant inputs rules out any straightforward use of Theorem 3.1. Hence, a systematic approach to time-scale decomposition for the system (4.1)–(4.4) must be developed.

As pointed out in [2], each step in the input will excite the fast dynamics. For this reason, a family of slow manifolds will be considered, each of interest over one control interval.

Assumption 4.1: For a fixed $n \geq N$, there exist fixed positive scalars r_x , r_z and r_u such that $u(t) \in \mathcal{B}_M(r_u)$ and $(x(t), z(t)) \in \mathcal{B}_N(r_x) \times \mathcal{B}_M(r_z)$ for all $t \in [0, nT]$, and such that $A_2(x)$ is nonsingular for all $x \in \mathcal{B}_N(r_x)$.

The family of slow manifolds is determined by setting $\epsilon = 0$ in (4.2), leading to

$$\phi_n(x_s(t), u[n]) := -A_2^{-1}(x_s(t))(a_2(x_s(t)) + B_2(x_s(t))u[n]). \quad (4.5)$$

The slow subsystem, which models the system assuming that $z(t) = \phi_n(x_s(t), u[n])$, is defined by

$$\frac{dx_s}{dt} = a_s(x_s(t)) + B_s(x_s(t))u[n], x_s(0) = x_0 \quad (4.6)$$

$$y_s(t) = c_s(x_s(t)) + D_s(x_s(t))u[n] \quad (4.7)$$

where

$$a_s(x_s) := a_1(x_s) - A_1(x_s)A_2^{-1}(x_s)a_2(x_s) \quad (4.8)$$

$$B_s(x_s) := B_1(x_s) - A_1(x_s)A_2^{-1}(x_s)B_2(x_s) \quad (4.9)$$

$$c_s(x_s) := c(x_s) - C(x_s)A_2^{-1}(x_s)a_2(x_s) \quad (4.10)$$

$$D_s(x_s) := -C(x_s)A_2^{-1}(x_s)B_2(x_s). \quad (4.11)$$

A unique solution $x_s(t)$ of the system (4.6)–(4.7) is assumed to exist for all $t \in [0, nT]$.

The family of fast subsystems, which models the deviation of z from its quasi-steady-state $\phi_n(x_s(nT), u[n])$, is defined by

$$\frac{dz_{f(n)}}{d\tau_n} = A_2(x_s(nT))z_{f(n)}(\tau_n) \quad (4.12)$$

$$y_{f(n)}(\tau_n) = C'(x_s(nT))z_{f(n)}(\tau_n) \quad (4.13)$$

in the fast time-scale

$$\tau_n := \frac{t - nT}{\epsilon}. \quad (4.14)$$

Each member could be initialized with $z(nT) = \phi_n(x(nT), u[n])$, but this approach will be taken only for $n = 0$. For $n \geq 1$, we substitute $x_s(nT)$ for $x(nT)$ and use the continuity of z at $t = nT$ to justify substitution of $z_{f(n-1)}(T/\epsilon) + \phi_{n-1}(x_s(nT), u[n-1])$ for $z(nT)$. Hence, the family of initial conditions is defined by

$$z_{f(n)}(0) = \begin{cases} z_0 - \phi_0(x_0, u[0]), n = 0 \\ z_{f(n-1)}(T/\epsilon) + \phi_{n-1}(x_s(nT), u[n-1]) \\ - \phi_n(x_s(nT), u[n]), n \geq 1. \end{cases} \quad (4.15)$$

Note that (4.12)–(4.15) are evaluated in a sequential nature, with the n th member of the family depending on the terminal value of the $(n-1)$ st member and on $x_s(nT)$. This formulation was motivated by (and encompasses) the resetting condition given in [9]. It is significant that (4.12)–(4.15), in general, depend on $x_s(nT)$.

The validity of the reduced-order subsystems as asymptotic approximations requires a stability assumption on (4.1)–(4.4).

Assumption 4.2: The fast states are asymptotically stable, i.e. $\text{Re}\{\lambda(A_2(x))\} \leq -\kappa < 0$ evaluated on $x = x_s$, for some fixed constant $\kappa > 0$.

The following theorem provides a method of approximating $x(t)$, $z(t)$ and $y(t)$ by solving the reduced-order subsystems (4.6)–(4.15).

Theorem 4.1: If the system (4.1)–(4.4) satisfies Assumptions 4.1 and 4.2, then there exists an $\epsilon^* > 0$ such that for any $\epsilon \in (0, \epsilon^*)$

$$x(t) = x_s(t) + O(\epsilon) \quad (4.16)$$

$$z(t) = \phi_n(x_s(t), u[n]) + z_{f(n)}(\tau_n) + O(\epsilon) \quad (4.17)$$

$$y(t) = y_s(t) + y_{f(n)}(\tau_n) + O(\epsilon) \quad (4.18)$$

for all $t \in [nT, (n+1)T)$, and $n = 0, \dots, \bar{n}$. Moreover, given any $T > 0$, there exists an $\epsilon^{**} \in (0, \epsilon^*)$ such that for any $\epsilon \in (0, \epsilon^{**})$

$$z_{f(n)}(T/\epsilon) = O(\epsilon) \quad (4.19)$$

for all $n = 0, \dots, \bar{n}$.

Proof: Over each interval $t \in [nT, (n+1)T)$, the system (4.1)–(4.4) can be viewed as a special case of system (3.1)–(3.3) parameterized by $u[n]$. From this point of view, Assumptions 4.1 and 4.2 imply Assumptions 3.1 and 3.2. Furthermore, note that (4.6)–(4.7) are equivalent to (3.4)–(3.5) except for the initial conditions taken at $t = nT$, and note that (4.12)–(4.13) are equivalent to (3.6)–(3.7) except for the parameterizing values of x and the initial conditions taken at $t = nT$. For $n = 0$, Theorem 3.1 may be directly applied (there are no exceptions for this index), proving the validity of (4.16)–(4.18) over $t \in [0, T)$. For $n \geq 1$, the exceptions can be summarized by noting that initial conditions (and parameterizing values of x) are assigned according to

$$x_s(nT) = \begin{cases} x(nT), \S 3.1 \\ x_s(nT), \S 4.1 \end{cases} \quad (4.20)$$

$$z_{f(n)}(0) = \begin{cases} z(nT) - \phi_n(x(nT), u[n]), \S 3.1 \\ z_{f(n-1)}\left(\frac{T}{\epsilon}\right) + \phi_{n-1}(x_s(nT), u[n-1]), \S 4.1 \\ -\phi_n(x_s(nT), u[n]), \S 4.1. \end{cases} \quad (4.21)$$

Now suppose that (4.16)–(4.18) hold for some $n' \geq 0$. Since x and z are continuous, $x((n'+1)T) = x_s((n'+1)T) + O(\epsilon)$ and $z((n'+1)T) = \phi_{n'}(x_s((n'+1)T), u[n']) + z_{f(n')}(T/\epsilon) + O(\epsilon)$. Hence, the discrepancies between the assignments indicated by (4.20)–(4.21) at $n = n' + 1$ are $O(\epsilon)$. Since all the differential equations involved admit solutions which depend continuously on their initial conditions (and parameterizing values of x), we see that (4.16)–(4.18) hold for $n' + 1$. By induction, (4.16)–(4.18) are proven for all $n = 0, \dots, \bar{n}$. The last claim, (4.19), is an easy consequence of Assumption 4.2. \square

B. Synthesis of Reduced-Order Observers with Inputs

In order to design reduced-order observers, discrete-time representations of the subsystem models are needed. The slow subsystem is sampled at the instants $t = nT$ to obtain

$$x_s[n+1] = F_s(x_s[n], u[n]) \quad (4.22)$$

$$y_s[n] = h_s(x_s[n], u[n]) \quad (4.23)$$

where $x_s[n] := x_s(nT)$, $y_s[n] := y_s(nT)$ and

$$h_s(x_s[n], u[n]) := c_s(x_s[n]) + D_s(x_s[n])u[n]. \quad (4.24)$$

The function F_s is given by [10]

$$F_s(x_s[n], u[n]) := \sum_{i=0}^{\infty} \frac{T^i}{i!} \left[\left(L_{a_s} + \sum_{j=1}^P u_j[n] L_{b_{s,j}} \right)^i x_s \right]_{x_s = x_s[n]} \quad (4.25)$$

where $b_{s,j}$ is the j th column of B_s , and u_j is the j th element of u . The family of fast subsystems is sampled at the instants $\tau_n = mT/(k\epsilon)$, where k is an integer ($k \geq M$) such that $T/k = O(\epsilon)$, to obtain

$$z_{f(n)}[m+1] = \Phi_{f(n)} z_{f(n)}[m] \quad (4.26)$$

$$y_{f(n)}[m] = C_{f(n)} z_{f(n)}[m] \quad (4.27)$$

where $z_{f(n)}[m] := z_{f(n)}(mT/(k\epsilon))$, $y_{f(n)}[m] := y_{f(n)}(mT/(k\epsilon))$ and

$$\Phi_{f(n)} := \exp \left(A_2(x_s[n]) \frac{T}{k\epsilon} \right) \quad (4.28)$$

$$C_{f(n)} := C(x_s[n]). \quad (4.29)$$

Note that the fast index m is implicitly associated with a slow index n . For simplicity, this dependence is suppressed in the notation. Note also that throughout this section, time-varying quantities are written in a manner consistent with the notation of Theorem 4.1.

The reduced-order observers are constructed by formulating the appropriate state-to-measurement maps. For each n , the incremental fast state-to-measurement map

$$\Delta H_{f(n)}(z_{f(n)}) := \begin{bmatrix} C_{f(n)}(\Phi_{f(n)} - I_M)z_{f(n)} \\ C_{f(n)}\Phi_{f(n)}(\Phi_{f(n)} - I_M)z_{f(n)} \\ \vdots \\ C_{f(n)}\Phi_{f(n)}^{M-1}(\Phi_{f(n)} - I_M)z_{f(n)} \end{bmatrix} \quad (4.30)$$

is implicitly parameterized by $x_s[n]$, and satisfies the identity

$$\Delta Y_{f(n)}[m] = \Delta H_{f(n)}(z_{f(n)}[m-M]) \quad (4.31)$$

where $\Delta Y_{f(n)}[m]$ is the block of incremental fast subsystem outputs

$$\Delta Y_{f(n)}[m] := \begin{bmatrix} y_{f(n)}[m-M+1] - y_{f(n)}[m-M] \\ \vdots \\ y_{f(n)}[m] - y_{f(n)}[m-1] \end{bmatrix}. \quad (4.32)$$

Assumption 4.3: For some positive scalars r_s and δ_f , there exist functions $\Delta H_{f(n)}^{-1} : \mathcal{R}^{QM} \rightarrow \mathcal{R}^M$ for $n = 0, \dots, \bar{n}$, parameterized by $x_s[n]$, such that

- 1) $\Delta Y_{f(n)} = \Delta H_{f(n)}(z_{f(n)})$ implies $z_{f(n)} = \Delta H_{f(n)}^{-1}(\Delta Y_{f(n)})$ for all $x_s[n] \in \mathcal{B}_N(r_s)$.
- 2) $\Delta H_{f(n)}^{-1}$ are continuous in the parameterizing $x_s[n]$ over the set $\mathcal{B}_N(r_s + \delta_f)$.

From (4.30), $\Delta H_{f(n)}^{-1}$ is a matrix pseudo-inverse. However, since $\Delta H_{f(n)}$ depends on x_s , the inversion may not in general be done off-line. Using Assumption 4.3, if $x_s[n] \in \mathcal{B}_N(r_s)$, then

$$z_{f(n)}[m] = \Phi_{f(n)}^M \Delta H_{f(n)}^{-1} (\Delta Y_{f(n)}[m]) \quad (4.33)$$

which is computable for $m = M, \dots, k$ and $n = 0, \dots, n$.

In constructing the slow observer, it will be convenient to use the notation $h_s^u(x_s) := h_s(x_s, u)$ and $F_s^u(x_s) := F_s(x_s, u)$. Summing slow and fast subsystem output blocks yields the identity

$$Y_{sf}[n] = \begin{bmatrix} h_s^{u[n-N+1]}(x_s[n-N+1]) \\ h_s^{u[n-N+2]} \circ F_s^{u[n-N+1]}(x_s[n-N+1]) \\ \vdots \\ h_s^{u[n]} \circ F_s^{u[n-1]} \circ \dots \circ F_s^{u[n-N+1]}(x_s[n-N+1]) \end{bmatrix} + \begin{bmatrix} C(x_s[n-N+1])z_{f(n-N+1)}[0] \\ C \circ F_s^{u[n-N+1]}(x_s[n-N+1])z_{f(n-N+2)}[0] \\ \vdots \\ C \circ F_s^{u[n-1]} \circ \dots \circ F_s^{u[n-N+1]}(x_s[n-N+1])z_{f(n)}[0] \end{bmatrix} \quad (4.34)$$

where

$$Y_{sf}[n] := \begin{bmatrix} y_s[n-N+1] \\ \vdots \\ y_s[n] \end{bmatrix} + \begin{bmatrix} y_{f(n-N+1)}[0] \\ \vdots \\ y_{f(n)}[0] \end{bmatrix}. \quad (4.35)$$

The slow-plus-fast state-to-measurement map² is defined as

$$H_{sf}(x_s, Z_f[n-1], U[n-1]) := \begin{bmatrix} h_s^{u[n-N]}(x_s) \\ h_s^{u[n-N+1]} \circ F_s^{u[n-N+1]}(x_s) \\ \vdots \\ h_s^{u[n-1]} \circ F_s^{u[n-1]} \circ \dots \circ F_s^{u[n-N+1]}(x_s) \end{bmatrix} + \begin{bmatrix} C(x_s)z_{f(n-N)}[k] \\ C \circ F_s^{u[n-N+1]}(x_s)z_{f(n-N+1)}[k] \\ \vdots \\ C \circ F_s^{u[n-1]} \circ \dots \circ F_s^{u[n-N+1]}(x_s)z_{f(n-1)}[k] \end{bmatrix} \quad (4.36)$$

which is parameterized by the vectors

$$Z_f[n-1] := \begin{bmatrix} z_{f(n-N)}[k] \\ \vdots \\ z_{f(n-1)}[k] \end{bmatrix} \quad (4.37)$$

$$U[n-1] := \begin{bmatrix} u[n-N] \\ \vdots \\ u[n-1] \end{bmatrix}. \quad (4.38)$$

Note that the right-hand side of (4.34) is not the same as the right-hand side of (4.36). However, substitution of the initial

²Since the fast output block in (4.34) depends on x_s , an implementable slow observer cannot be formulated using a "slow" state-to-measurement map, in contrast to the previous section.

conditions (4.15) (along with the definition (4.24)) into (4.34) proves that this map satisfies

$$H_{sf}(x_s[n-N+1], Z_f[n-1], U[n-1]) = Y_{sf}[n]. \quad (4.39)$$

Assumption 4.4. For the fixed positive scalar r_s and some positive scalars r_f and δ_s , there exists a function $H_{sf}^{-1} : \mathcal{R}^{QN} \times \mathcal{R}^{MN} \times \mathcal{R}^{PN} \rightarrow \mathcal{R}^N$ such that

- 1) $x_s \in \mathcal{B}_N(r_s)$ and $Y_{sf} = H_{sf}(x_s, Z_f, U)$ imply $x_s = H_{sf}^{-1}(Y_{sf}, Z_f, U)$ for all $(Z_f, U) \in \mathcal{B}_{MN}(r_f) \times \mathcal{B}_{PN}(r_u)$.
- 2) $H_{sf}^{-1}(\cdot, \cdot, U)$ is continuous on the set $\mathcal{B}_{QN+MN}(r + \delta_s)$ where $r > 0$ is such that $H_{sf}(\mathcal{B}_N(r_s), \mathcal{B}_{MN}(r_f), \mathcal{B}_{PN}(r_u)) \times \mathcal{B}_{MN}(r_f) \subseteq \mathcal{B}_{QN+MN}(r)$.

Using Assumption 4.4, if $x_s[n-N+1] \in \mathcal{B}_N(r_s)$ and $Z_f[n] \in \mathcal{B}_{MN}(r_f)$ then

$$x_s[n] = F_s^{u[n-1]} \circ \dots \circ F_s^{u[n-N+1]} \circ H_{sf}^{-1}(Y_{sf}[n], Z_f[n-1], U[n-1]) \quad (4.40)$$

which is computable for $n = N, \dots, n$.

Neither (4.33) nor (4.40) is implementable. The unknown signals are estimated by

$$\Delta \hat{Y}_{f(n)}[m] := \begin{bmatrix} y(nT' + (m-M+1)T/k) - y(nT' + (m-M)T/k) \\ \vdots \\ y(nT' + mT/k) - y(nT' + (m-1)T/k) \end{bmatrix} \quad (4.41)$$

$$Y_{sf}[n] := \begin{bmatrix} y((n-N+1)T) \\ \vdots \\ y(nT) \end{bmatrix} \quad (4.42)$$

$$Z_f[n-1] := \begin{bmatrix} z_{f(n-N)}[k] \\ \vdots \\ z_{f(n-1)}[k] \end{bmatrix} \quad (4.43)$$

where $\hat{z}_{f(n)}[m]$ will be supplied by the fast observer. Also, $\Phi_{f(n)}$ and $\Delta H_{f(n)}^{-1}$ require knowledge of $x_s[n]$, motivating the following assumption.

Assumption 4.5. At least one of the following conditions is satisfied:

- 1) $\hat{x}_0 = x_0 + O(\epsilon)$ is known.
- 2) The matrices $A_2(x) = A_2$ and $C(x) = C$ are constant.

If Assumption 4.5-2 is true, then both $\Phi_{f(n)}$ and $\Delta H_{f(n)}^{-1}$ are independent of $x_s[n]$. If Assumption 4.5-1 is true, then $\Phi_{f(n)}$ and $\Delta H_{f(n)}^{-1}$ are estimated according to

$$\hat{\Phi}_{f(n)} := \Phi_{f(n)}|_{x_s[n] = \hat{x}_s[n]} \quad (4.44)$$

$$\Delta \hat{H}_{f(n)}^{-1}(z_{f(n)}) := \Delta H_{f(n)}^{-1}(z_{f(n)})|_{x_s[n] = \hat{x}_s[n]} \quad (4.45)$$

where

$$\hat{x}_s[n] := \begin{cases} \hat{x}_0, n=0 \\ F_s^{u[n-1]} \circ \dots \circ F_s^{u[0]}(\hat{x}_0), n=1, \dots, N-1 \end{cases} \quad (4.46)$$

and where $\hat{x}_s[n]$ will be supplied by the slow observer for $n = N, \dots, \bar{n}$.

The implementable reduced-order observers are defined in the following theorem, and they are interconnected in the sense that $\hat{x}_s[n]$ is used to compute $\hat{z}_{f(n)}[k]$, which in turn is used to compute $\hat{x}_s[n+1]$, and so on.

Theorem 4.2: Suppose $(x_s[n], z_{f(n)}[m]) \in \mathcal{B}_N(r_s) \times \mathcal{B}_M(r_f)$ for $n = 0, \dots, \bar{n}$ and $m = 0, \dots, k$. If the system (4.1)–(4.4) satisfies Assumptions 4.1–4.5, then there exists an $\epsilon^* > 0$ such that for any $\epsilon \in (0, \epsilon^*)$

$$\hat{z}_{f(i)}[m] := \hat{\Phi}_{f(i)}^M \Delta \hat{H}_{f(i)}^{-1} (\Delta \hat{Y}_{f(i)}[m]) \quad (4.47)$$

$$= z_{f(i)}(mT/(k\epsilon)) + O(\epsilon) \quad (4.48)$$

$$\hat{x}_s[n] := F_s^{u[n-1]} \circ \dots \circ F_s^{u[n-N+1]} \circ H_{sf}^{-1} (\hat{Y}_{sf}[n], \hat{Z}_f[n-1], U[n-1]) \quad (4.49)$$

$$= x_s(nT) + O(\epsilon) \quad (4.50)$$

for all $i = 0, \dots, \bar{n}$, $n = N, \dots, \bar{n}$, and $m = M, \dots, k$.

Proof: By (4.6) and the fact that $T/k = O(\epsilon)$, it is evident that $x_s(iT + mT/k) = x_s(iT + (m+1)T/k) + O(\epsilon)$. The continuity of h_s , along with (4.18) implies that

$$\Delta \hat{Y}_{f(i)}[m] = \Delta Y_{f(i)}[m] + O(\epsilon). \quad (4.51)$$

This fact, along with Assumptions 4.3 and 4.5 and (4.46), implies (4.48) for $i = 0, \dots, N-1$. We now proceed inductively in n . The above analysis proves that

$$\hat{Z}_f[n'-1] = Z_f[n'-1] + O(\epsilon) \quad (4.52)$$

for $n' = N$. Also, (4.18) implies that

$$\hat{Y}_{sf}[n'] = Y_{sf}[n'] + O(\epsilon) \quad (4.53)$$

for $n' = N$. These two facts, along with Assumption 4.4, the continuity of F_s and the hypothesis that $(x_s[n'], Z_f[n'-1]) \in \mathcal{B}_N(r_s) \times \mathcal{B}_{MN}(r_f)$ imply (4.50) for $n = n' = N$. But $\hat{x}_s[n'] = x_s[n'] + O(\epsilon)$ in turn implies (as argued above) that $\hat{z}_{f(n')}[m] = z_{f(n')}[m] + O(\epsilon)$ for $m = M, \dots, k$. Thus, repetition of this argument at an arbitrary n' inductively proves the result. Note finally that since this induction stops when $n' = n$, any cumulative errors are guaranteed to be $O(\epsilon)$. \square

Since the observers (4.47) and (4.49) are quite complex, it is helpful to enumerate the sequence of operations that implements the algorithm. The observers are evaluated by:

- 1) Using $\hat{x}_s[0] = \hat{x}_0$ if necessary, construct (4.44) and (4.45) for $n = 0$.
- 2) Set $m = n = 0$.
- 3) While $n \leq \bar{n}$
 - a) Measure $y(nT + mT/k)$.
 - b) If $m \geq M$, compute $\hat{z}_{f(n)}[m]$ from (4.47) using $y(nT + (m-M)T/k), \dots, y(nT + mT/k)$.
 - c) If $m = k$, then
 - Set $n = n + 1$ and $m = 0$.
 - If $n \geq N$, compute $\hat{x}_s[n]$ from (4.49) using $y((n-N+1)T), \dots, y(nT)$, $\hat{z}_{f(n-N)}[k], \dots, \hat{z}_{f(n-1)}[k]$ and $u[n-N], \dots, u[n-1]$. Else, set $\hat{x}_s[n]$ according to (4.46).

- Using $\hat{x}_s[n]$ if necessary, construct (4.44) and (4.45).

d) Set $m = m + 1$.

4) Stop.

Under certain conditions, it is possible to construct a slow observer which is independent of the fast observer. This is possible because if $z_{f(n)}$ is fast enough relative to T , then its contribution to the slow observer may be neglected.

Corollary 4.1: Suppose $x_s[n] \in \mathcal{B}_N(r_s)$ for $n = 0, \dots, \bar{n}$. If the system (4.1)–(4.4) satisfies Assumptions 4.1, 4.2, and 4.4, then there exists an $\epsilon^{**} > 0$ such that for any $\epsilon \in (0, \epsilon^{**})$

$$\tilde{x}_s[n] := F_s^{u[n-1]} \circ \dots \circ F_s^{u[n-N+1]} \circ H_{sf}^{-1} (\hat{Y}_{sf}[n], 0, U[n-1]) \quad (4.54)$$

$$= x_s(nT) + O(\epsilon) \quad (4.55)$$

for $n = N, \dots, \bar{n}$.

Proof: In light of (4.19), there exists an ϵ^{**} so that (for the given T) $z_{f(n)}[k] = O(\epsilon)$ for all $n = 0, \dots, \bar{n}$. Utilizing this fact in (4.49) yields (4.55). \square

Corollary 4.1 is different from Corollary 3.1 in that (4.54) produces estimates over the same range of n as (4.49) does. This is because ϵ is assumed to be small enough so that the boundary layers are over in one slow sampling interval. Note that it is not possible to simply “wait long enough” for the boundary layer to be over, because it is excited by each step in the slow input. Finally, the observer (4.54) does not require either fast subsystem observability, knowledge of x_0 , or special structure of A_2 and C .

C. Electromechanical System Example

An important practical application considered now is velocity control of two-phase permanent-magnet synchronous motors using only phase current measurements. In [6], a recursive nonlinear observer similar to an Extended Kalman Filter is constructed to estimate the rotor motion from stator current measurements. The disadvantages of the scheme in [6], in comparison with the methods introduced in this paper, include: sampling and computation necessarily occur in the electrical time-scale; the innovation gains must be selected for a specific and limited range of rotor velocities to avoid poor estimation error convergence.

The two-phase permanent-magnet synchronous motor can indeed be modeled as a singularly perturbed nonlinear system in the form (4.1)–(4.4) under standard assumptions. Assuming that magnetic saturation effects are negligible, that the air gap is smooth, that the distribution of windings eliminates harmonic effects, and that the load is a constant inertia, the continuous-time dynamic model is

$$\frac{d\theta}{dt} = \omega \quad (4.56)$$

$$\frac{d\omega}{dt} = \frac{1}{J} (K_m (i_1 \sin(N_p \theta) + i_2 \cos(N_p \theta)) - B\omega) \quad (4.57)$$

$$L \frac{di_1}{dt} = v_1 - Ri_1 - K_m \omega \sin(N_p \theta) \quad (4.58)$$

$$L \frac{di_2}{dt} = v_2 - Ri_2 - K_m \omega \cos(N_p \theta) \quad (4.59)$$

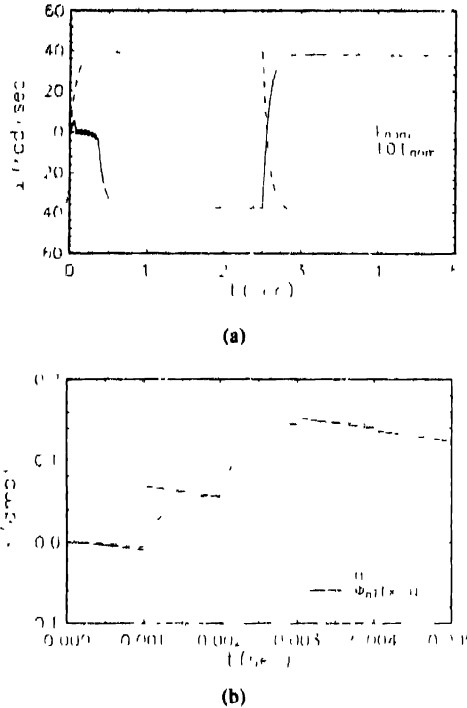


Fig. 3. Motor response under output feedback control using only the slow observer of Corollary 4.1 for $L = L_{nom}$ and $L = 10L_{nom}$. (a) Actual rotor velocity $\omega(t)$ and (b) Phase 1 current $i_1(t)$ and slow manifolds $\phi_n(x_s(t), u[n])$ for $n = 0, \dots, 4$ at $L = L_{nom}$.

where θ and ω are the angular position and velocity of the rotor, i_1 and i_2 are the currents in the stator phases, v_1 and v_2 are the voltages applied to the stator phases, K_m is the magnet constant, N_p is the number of magnet pole pairs on the rotor, J is the total rotor inertia, B is the total rotor viscous damping coefficient, L is the inductance of the stator phases, and R is the resistance of the stator phases. Typically, L can be taken as a small parameter so that, in the notation of (4.1)–(4.3), $\epsilon = L$, $x = [\theta \ \omega]'$, $z = [i_1 \ i_2]'$, $u = [v_1 \ v_2]'$, and $y = [i_1 \ i_2]'$.

The reduced-order observers are formulated according to Theorem 4.2. The fast observer is decoupled from the slow observer and is defined for all $1 \leq m \leq k$ and each fixed n by

$$\hat{i}_{f(n)1}[m] = \frac{\phi_f}{\phi_f - 1} (i_{n1}[m] - i_{n1}[m-1]) \quad (4.60)$$

$$\hat{i}_{f(n)2}[m] = \frac{\phi_f}{\phi_f - 1} (i_{n2}[m] - i_{n2}[m-1]) \quad (4.61)$$

where $i_n[m] := i(nT + mT/k)$ and $\phi_f := \exp(-RT/Lk)$. The slow observer can be constructed using only the slow subsystem output equations. Thus, discretization of the slow subsystem is not necessary, and the slow observer is implicitly defined for $n \geq 1$ as a solution $(\hat{\theta}_s[n], \hat{\omega}_s[n])$ of the system of nonlinear algebraic equations

$$\begin{aligned} \psi_1 &:= \frac{1}{K_m} (v_1[n-1] - R(i_1[n] - \hat{i}_{f(n-1)1}[k])) \\ &= \hat{\omega}_s[n] \sin(N_p \hat{\theta}_s[n]) \end{aligned} \quad (4.62)$$

$$\begin{aligned} \psi_2 &:= \frac{1}{K_m} (v_2[n-1] - R(i_2[n] - \hat{i}_{f(n-1)2}[k])) \\ &= \hat{\omega}_s[n] \cos(N_p \hat{\theta}_s[n]) \end{aligned} \quad (4.63)$$

where $i[n] := i(nT)$, and where ψ_1 and ψ_2 are computed from known quantities. The observer of Corollary 4.1 is recovered by setting $\hat{i}_{f(n-1)}[k] = 0$ in (4.62)–(4.63).

A few comments are in order concerning the inversion of the slow observer equations. Whenever $\psi_1 \neq 0$ or $\psi_2 \neq 0$, (4.62)–(4.63) can be solved for $\hat{\theta}_s[n]$ and $\hat{\omega}_s[n]$. There are, however, multiple solutions. If (θ^0, ω^0) is a solution, then so are $(\theta^0 \pm \frac{2\pi}{N_p}j, \omega^0)$ and $(\theta^0 \pm \frac{\pi}{N_p}(2j-1), -\omega^0)$, for any integer j . In the previous section, this type of difficulty was avoided by insisting that $x_s[n] \in \mathcal{B}_N(r_s)$. In this example, such an approach would require $\theta_s[n] \in \mathcal{B}_1(\pi/2N_p - \delta_\theta)$ for some $\delta_\theta \in (0, \pi/2N_p)$. Since limiting the rotor motion is not practical, heuristic rules will be used to decide which solution of (4.62)–(4.63) is the correct one. Assuming that T is sufficiently small so that θ does not vary more than $\pm\pi/(2N_p)$ from one slow sampling instant to the next, it is reasonable to choose the solution closest to the previous estimate. If the motor starts from rest and $\theta(0)$ is known to $\pm\pi/(2N_p)$, then this scheme will provide the correct solution. If $\psi_1 = \psi_2 = 0$, then $\hat{\omega}_s[n] = 0$ and thus (4.62)–(4.63) cannot be inverted, so the estimate $\hat{\theta}_s[n] = \hat{\theta}_s[n-1]$ will be used in this case.

The observer-based feedback will be designed from the slow subsystem. The simplest such feedback, one that emulates a continuous-time linearizing slow controller, is given by

$$\begin{aligned} \begin{bmatrix} v_1[n] \\ v_2[n] \end{bmatrix} &= \begin{bmatrix} \sin(N_p \hat{\theta}_s[n]) \\ \cos(N_p \hat{\theta}_s[n]) \end{bmatrix} \\ &\times \left(\frac{K_m^2 + BR - JRK_\omega}{K_m} \omega_s[n] + \frac{JRK_\omega}{K_m} \omega^d \right) \end{aligned} \quad (4.64)$$

where ω^d is the desired velocity and $K_\omega > 0$ is a design parameter. The continuous-time version of (4.64) acting on the slow subsystem (under the assumptions that $L = 0$, $\theta_s = \theta$, and $\hat{\omega}_s = \omega_s$) results in closed-loop dynamics $\dot{\theta}_s = \omega_s$, $\dot{\omega}_s = K_\omega(\omega^d - \omega_s)$. Since (4.64) emulates a continuous-time control, T needs to be relatively small.

Simulations were run with parameters from a commercially available motor: $N_p = 2$ pole pairs, $R = 0.418\Omega$, $L = 0.131$ mH, $K_m = 29.4$ mN·m/A, $J = 0.0205$ g·m²/rad, $B = 0.013$ g·m²/rad·sec, $T = 0.001$ sec and $k = 10$. Note that the value given above for L (i.e. ϵ) will be used as a “nominal” value. In all cases, the motor, starting from rest with $\hat{\theta}_s[0] = 0$ and $\theta(0) = 0.2$, was commanded to regulate to $\omega^d = 40$ rad/sec for $t \in [0, 2.5]$ sec, and to $\omega^d = -40$ rad/sec for $t \in [2.5, 5.0]$ sec, with $K_\omega = 13$.

The first case considered corresponds to output feedback control using the slow observer of Corollary 4.1. Fig. 3(a) shows $\omega(t)$ for $L = L_{nom}$ and $L = 10L_{nom}$. At $L = L_{nom}$, the controller does an excellent job of providing the desired linear response. At $L = 10L_{nom}$, however, the controller has trouble at startup, and the heuristics for choosing the correct sign of the velocity fail. In Fig. 3(b), $i_1(t)$ and the corresponding first five slow manifolds are shown for $L = L_{nom}$. In this case, T is just large enough for $i_1(t)$ to settle to its quasi-steady-state value. For $L = 10L_{nom}$, this property no longer holds and, hence, the observer fails severely.

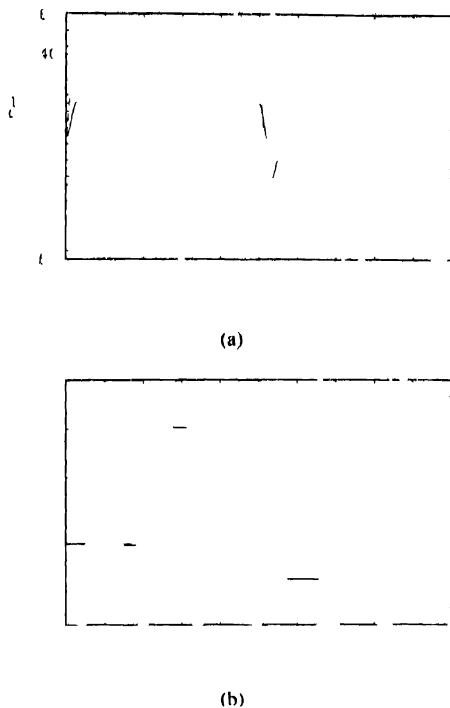


Fig. 4 Motor response under output feedback control using both the slow and fast observers of Theorem 4.2 for $L = 10I_{nom}$ and $L = 500I_{nom}$. (a) Actual rotor velocity $\omega(t)$ and (b) Phase 1 current $i_1(t)$ and slow manifolds $u[n]$ for $n = 0, \dots, 1$ at $L = 10I$.

The second case considered corresponds to output feedback control using the slow and fast observers of Theorem 4.2. Fig. 4(a) shows $\omega(t)$ for $L = 10I_{nom}$ and $L = 500I_{nom}$. At $L = 10I_{nom}$ the controller performs very well although a perceptible steady-state error occurs. Although not evident from the plot the steady state error is not a consequence of observation error but instead is due to the failure of the emulating controller. At $L = 500I_{nom}$ the controller yields substantially degraded performance. In Fig. 4(b), $i_1(t)$ and the corresponding first five slow manifolds are shown for $L = 10I_{nom}$. Even though $i_1(t)$ doesn't settle to its quasi-steady-state value, use of the fast observer has accounted for this feature, resulting in good performance.

V CONCLUSIONS

In this paper, the observer design problem for nonlinear singularly perturbed systems has been formulated as one of designing separate observers for the approximate slow and fast subsystem models. Not only are the resulting observers of reduced-order, they are also implemented in a multirate fashion. Both of these features result in significant savings in computational and/or memory requirements. Furthermore, the stiffness inherent in a full-order implementation is eliminated by this approach.

It has been proven that this reduced-order observer design methodology guarantees $O(\epsilon)$ observation errors. Hence for sufficiently small ϵ , these errors will be acceptable for all practical purposes. Although the development of stabilizing observer-based control algorithms has not been attempted, the use of observer-based feedback has been considered in a case-

study. For the application to so-called sensorless control of permanent-magnet synchronous motor, the performance has been demonstrated using our reduced-order observer designs.

REFERENCES

- [1] J. P. Barbot, S. Monaco, D. Normand-Cyrot, and N. Poullos, "Discretization schemes for nonlinear singularly perturbed systems," in *Proc. 30th IEEE Conf. Decis. Contr.*, Brighton, England, Dec. 1993, pp. 443-448.
- [2] F. Esfandiari and H. K. Khalil, "On the robustness of sampled data control to unmodelled high frequency dynamics," *IEEE Trans. Automat. Contr.*, vol. 34, no. 8, pp. 900-903, 1989.
- [3] J. W. Grizzle and P. V. Kokotovic, "Feedback linearization of sampled data systems," *IEEE Trans. Automat. Contr.*, vol. AC-33, no. 9, pp. 857-859, 1988.
- [4] J. W. Grizzle and P. T. Moraal, "On Observers for Smooth Nonlinear Digital Systems," in *Lecture Notes in Control and Information Sciences*, vol. 144, 1990, pp. 401-410.
- [5] ———, "Newton observers and nonlinear discrete time control," in *Proc. 29th IEEE Conf. Decis. Contr.*, Honolulu, Hawaii, Dec. 1990, pp. 760-767.
- [6] L. A. Jones and J. H. Lang, "A state observer for the permanent magnet synchronous motor," *IEEE Trans. Indust. Electron.*, vol. 36, no. 3, pp. 374-382, 1989.
- [7] H. K. Khalil, *Nonlinear Systems*, New York: Macmillan, 1992.
- [8] P. V. Kokotovic, H. K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control Analysis and Design*, New York: Academic, 1986.
- [9] B. Litkouhi and H. K. Khalil, "Multirate and composite control of two time scale discrete time systems," *IEEE Trans. Automat. Contr.*, vol. AC-30, no. 7, pp. 645-651, 1985.
- [10] S. Monaco and D. Normand-Cyrot, "On the sampling of a linear analytic control system," in *Proc. 24th IEEE Conf. Decis. Contr.*, Fort Lauderdale, FL, Dec. 1985, pp. 1457-1462.



Kenneth R. Shouse (S'88, M'93) received the B.S. degree in mechanical engineering in 1986 from Texas A&M University, College Station. He received the M.S. and Ph.D. degrees in electrical engineering in 1991 and 1993, respectively, from the Georgia Institute of Technology, Atlanta.

From 1986-1988, Dr. Shouse worked as a Project Engineer for Viste Chemical Co. in Lake Charles, LA. He is currently a Research Engineer in the Engine and Vehicle Research Division of the Southwest Research Institute, San Antonio, TX. His research interests include discrete time nonlinear control theory and its application to internal combustion engines and electromechanical systems.



David G. Taylor (S'80, M'88) was born in Oak Ridge, Tennessee, on August 7, 1961. He received the B.S. degree from the University of Tennessee, Knoxville, in 1983 and the M.S. and Ph.D. degrees from the University of Illinois, Urbana-Champaign, in 1985 and 1988, respectively, all in electrical engineering.

During the summers from 1981 to 1983, he was employed by Oak Ridge National Laboratory, Oak Ridge, Tennessee, and IBM Corporation, Charlotte, North Carolina. From 1983 to 1988, he held various fellowships and assistantships at the University of Illinois, Urbana-Champaign. Since 1988, he has been with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, where he is now an Associate Professor.

Dr. Taylor's research interests include nonlinear control theory and its applications to electromechanical systems. He received the NSF Presidential Young Investigator Award in 1991.

Adaptive Back-Pressure Congestion Control Based on Local Information

Leandros Tassiulas, *Member, IEEE*

Abstract—The problem of distributed congestion control as it arises in communication networks as well as in manufacturing systems is studied in this paper. In particular, a multistage queueing system that models virtual circuit and datagram communication networks and a class of manufacturing systems are considered. The topology may be arbitrary, there are multiple traffic classes, and the routing can be class dependent, with routes that may form direct or indirect loops. The model incorporates the functions of transmission scheduling, flow control, and routing, through which congestion control is performed in the network. A policy is given that performs these functions jointly. According to the policy, heavily loaded queues are given higher priority in service. A congested node may reduce the flow from upstream nodes through a flow control mechanism. Whenever routing is required, it is performed in such a manner that the lightly loaded queues receive most of the traffic. For arrival processes with bounded burstiness, the policy guarantees bounded backlogs in the network, as long as the load of each server is less than one. The actions of each server are based on the state of its own queues and of the queues one hop away. Therefore, they are implementable in a distributed fashion. An adaptive version of the policy is also provided which makes it independent of the arrival rates.

I. INTRODUCTION

DYNAMIC control achieves a better utilization of the transmission and switching resources of a communication network over static schemes. Dynamic schemes have been adopted for packet routing, for sharing the transmission capacity of a link among several competing information streams, or for controlling the flow of traffic in a packet switching network. In large networks though, where the control functions are distributed, dynamic control policies are prone to misbehavior, and the network may enter intriguing instability modes. The same phenomenon has been observed in manufacturing systems with distributed scheduling [10], [6]. In this work, we present a new distributed dynamic control policy in a multistage queueing system that models virtual circuit and datagram communication networks, as well as a class of manufacturing systems. The policy, which we call the adaptive back-pressure (ABP) congestion control policy, schedules the servers, routes the served traffic, and controls the flow based on local information, and therefore is amenable to distributed implementation. Each server is allocated dynamically based on the state of the queues that it serves, as well as on the state

of downstream queues, which are one hop away. When there are routing options, the decision is taken again in a similar distributed fashion. It is shown that under the ABP policy, the backlog in the network remains bounded as long as the utilization of each server is less than one.

A central problem in the design of virtual circuit (VC) communication networks is the sharing of the transmission capacity of a link among the virtual circuits that go through it. Several schemes have been proposed, including round robin, weighted round robin, golden ratio scheduling, virtual clock, and processor sharing [4], [9], [14], [8], [5]. In some of the above schemes [4], [5], the link transmission scheduling is combined with window flow control. One of the main problems arising in this context is how to evaluate the throughput of a virtual circuit network, that is, the session rates that can be sustained by the network for certain transmission control policies, without experiencing instabilities. Cruz [2], [3] has considered this problem in virtual circuit networks by employing several different link transmission disciplines including FIFO, LIFO, and strict priority. He obtained bounds for the backlog in the nodes of the network. These bounds depend on the session arrival rates and burstiness characteristics, and guarantee stable operation of the system for the traffic parameters for which they remain finite. In certain cases though, in networks with virtual circuits that may form cycles, these bounds explode for arrival rates which give utilization strictly less than one at all network links. In this case, no conclusion can be drawn for the stability of the network. Chang [1] extended these results, obtaining backlog bounds in multiclass networks with routing. The stability problem yet remained open for certain cases in networks with cycles. Yaron and Sidi [13] addressed the same question with processes satisfying constraints on the tails of the backlog distribution. Hahne [4] has shown that with round robin scheduling in each link and hop-by-hop window flow control, there exist window sizes to stabilize the network as long as the link utilization is less than one. The selection of the window sizes depends on the arrival rates. By the ABP policy, stabilization of the network is achieved as long as the utilization of every link is less than one, and no knowledge of the arrival rates is required for its implementation. When the routes of the packets are not prespecified but only their final destinations are given, as is the case with datagram communication networks, then the ABP policy combines routing with the scheduling mechanisms described in the VC case to preserve bounded queue lengths.

A problem of stability, similar to the one discussed above, arises in manufacturing systems operated under distributed

Manuscript received December 22, 1992; revised May 15, 1994. Recommended by Associate Editor, P. Nain. This work was supported in part by the Center for Advanced Technology in Telecommunications, Polytechnic University, and by the NSF under Grant NCR-9211417.

The author is with the Department of Electrical Engineering, Polytechnic University, Brooklyn, NY 11201 USA.

IEEE Log Number 9407563.

scheduling policies. Perkins and Kumar [10] and Kumar and Seidman [6] have studied the problem of stability in a flexible manufacturing system with distributed scheduling. While simple distributed policies are shown to stabilize acyclic manufacturing networks in [10], a simple two-stage queueing network is presented in [6]. In this model instabilities occur in situations in which all servers of the system are strictly underloaded and the system is operated under a work-conserving policy. This example demonstrates how instabilities may occur in multistage manufacturing systems due to the phenomena of starvation and overloading. The queueing network that corresponds to the manufacturing system considered falls within the scope of the networks presented in our study. The ABP policy stabilizes the manufacturing system as long as the utilization of each machine is less than one. The scheduling of each machine i is determined by the size of the backlog of the different part types in i , as well as the size of the backlogs of those part types in downstream machines one hop away from i . Occasionally, the machines are forced to idle again, depending on the local state. The queueing system presented here extends the above manufacturing systems model to include the case where a part type may have the option of several alternative manufacturing scenarios and routing decisions are made. In Section II-C, this is discussed in more detail. The ABP policy in that case combines scheduling and routing decisions to achieve the same goal. In [11], a single-class network was considered, and a routing policy of the same nature was studied for Poisson arrivals.

For the arrival process, we assume that the burstiness is bounded by a deterministic bound [2], [3]. This traffic assumption has been used widely lately since the outputs of the traffic regulators that shape the traffic before it enters a high-speed network satisfy this type of constraint. Lu and Kumar [7] have used a similar type of traffic in the context of a manufacturing system.

The rest of the paper is organized as follows. In Section II, the queueing network is presented, and the correspondence with virtual circuit and datagrams networks as well as manufacturing systems is demonstrated. In Section III, the issue of stability is discussed, and the sufficient stability condition is shown. In Section IV, a parametric class of policies is specified, and their stability is studied. In Section V, the ABP policy is specified and investigated. Finally, in Section VI, the results are discussed and some open problems are presented.

II. THE NETWORK MODEL

We consider a network consisting of N servers and B buffers (Fig. 1). Each server i can serve any buffer from the set of buffers B_i , $i = 1, \dots, N$ —it is allocated to the buffers according to some scheduling discipline. The sets B_i may be overlapping, that is, a buffer may be served by more than one server simultaneously. The served traffic from buffer j can be directed to any buffer of the set \mathcal{R}_j . A routing policy determines to which buffer in \mathcal{R}_j the traffic from j is routed. From certain buffers, the traffic can be directed out of the system; this is indicated by including a 0 in the set \mathcal{R}_j . We make the natural assumption that, from every buffer, the work

can be forwarded out of the system if it is routed through an appropriate sequence of buffers. We consider a fluid model for the work coming into the system. The time is continuous and the instantaneous rate with which work comes to buffer i from outside is $a_i(t)$. For simplicity, we assume that the arrival stream can contain no impulses, that is, the arrival rate $a_i(t)$ for all buffers i is bounded uniformly over time by a bound λ . In the time interval (t_1, t_2) , an amount of work equal to

$$\int_{t_1}^{t_2} a_i(t) dt$$

enters buffer i from outside. We assume that the arrival streams satisfy some burstiness constraints; that is, there are nonnegative numbers a_i, b_i , $i = 1, \dots, B$ such that, for all $0 \leq t_1 < t_2$, we have

$$\int_{t_1}^{t_2} a_i(t) dt \leq a_i(t_2 - t_1) + b_i. \quad (2.1)$$

We will assume that the long-run average rate with which work enters the system in buffer i exists

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t a_i(s) ds = a_i, \quad i = 1, \dots, B$$

and a_i will be referred to as the arrival rate to buffer i . The latter assumption is not needed for the validity of the results. It is introduced only because it is conceptually simpler to think of the a_i as being the arrival rate to buffer i . When server i provides service to buffer $j \in B_i$ which is nonempty, work leaves the buffer with a constant rate μ_{ij} , the service rate of server i at buffer j . If server i is assigned to buffer j continuously for an amount of time T and the traffic is directed to buffer l in \mathcal{R}_j , then an amount of work $T\mu_{ij}$ is transferred from j to l . The selection of buffer $j \in B_i$ served by i and of buffer $l \in \mathcal{R}_j$ to where the traffic is directed is done by the control policy. This work is routed to some buffer in \mathcal{R}_j . We allow more than one server to serve the same buffer; in that case, the service rate is assumed to be equal to the sum of the rates of the servers that serve the buffer. When server i that serves buffer j and directs the traffic to buffer k switches to buffers l and m , respectively, a switchover time δ_{iklm}^s is involved during which the server idles. The results in the paper hold for arbitrary and distinct switchover times. For notational simplification but no loss of generality, however, we will assume that the switchover time denoted by δ and be the same for all servers and buffers. In the following, we discuss how certain networks fall within the scope of the above model.

A. Virtual Circuit Networks

A VC network is characterized by its topology graph $G = (V, E)$, the transmission rates of the links and the established VC's. The topology graph contains one node for every network node and one directed link $e = (v, w)$ for every communication link from node v to node w . The transmission rate of link e is denoted by μ_e . A virtual circuit i is specified by the sequence of links e_1^i, \dots, e_N^i that traverses as it goes through the network. The traffic of VC i enters the network at the origin node of e_1^i and leaves it at the destination node

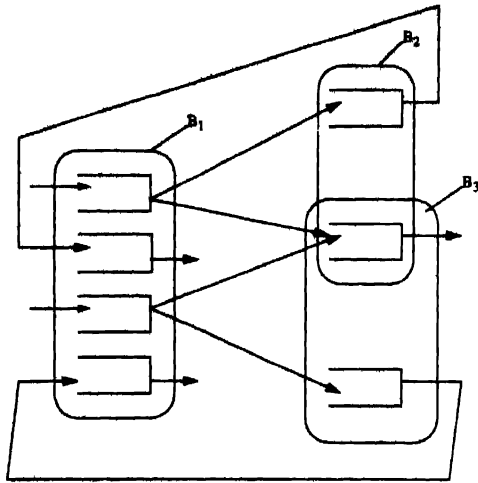


Fig. 1. An example of the network specified in Section II is depicted. There are three servers 1, 2, 3 (not depicted) which serve the queues in the sets B_1 , B_2 , and B_3 , respectively. The traffic from each queue i may be routed to any queue j to which there is an arrow from i .

of link e_N^i . A scheduling policy schedules for each link the transmissions of the virtual circuits going through it. Also, at certain times, the link may be forced to idle due to flow control actions.

A virtual circuit network is modeled by the above queueing system as follows (Fig. 2). Each link corresponds to a server with transmission capacity equal to its service rate. Clearly, in this case, the service rate of server (link) j is the same for all buffers in B_j . There is one buffer for every VC i and every link e_k^i is traversed by i . The buffers of virtual circuit i contain traffic of that VC waiting to be transmitted through the corresponding link. When a link is allocated to any of the buffers of the VC's going through it, that buffer empties with rate equal to the link transmission rate. The traffic from the buffer of virtual circuit i at link e_j^i is routed to the buffer of the same virtual circuit at link e_{j+1}^i , except if e_j^i is the last link traversed by VC i ; in the latter case, the traffic leaves the system. Hence, the set \mathcal{R}_j contains a single element for each j and no routing is needed. The traffic in the VC's consists of streams of packets, and clearly, the link cannot switch from VC to VC in a time period smaller than a packet transmission time. In the assumptions we made, though, in the previous section, the traffic is considered as a continuous flow. For the results that we obtain here, it is not important whether the traffic is continuous or in terms of packets since the policies we propose can be easily modified to work with packetized traffic.

B. Datagram Networks

In datagram networks, the traffic of a communication session does not have to follow a specific route, but can be routed arbitrarily, and different packets may follow different paths to the destination. The packets are differentiated only by their destination nodes. When a packet arrives at a node which is not its final destination, a decision is made as to which outgoing link the packet will follow next; then the packet is placed at the outgoing buffer of that link. There are no constraints on

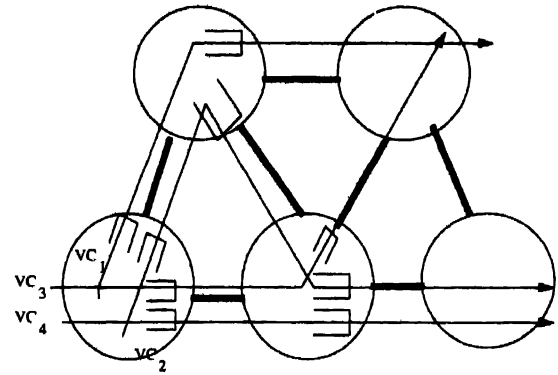


Fig. 2. A virtual circuit network

which outgoing link the packets of each specific destination will follow; this is determined by the routing policy. Packets of several different destinations are waiting in the buffer of each outgoing link to be transmitted. The link scheduling discipline determines which is the packet to be transmitted next. In the corresponding queueing model, the buffers are in one-to-one correspondence with the links. For each server i , there is one buffer in B_i for each destination node. The traffic with destination node k , which is transmitted through link i , is stored in the corresponding buffer of B_i . These buffers do not necessarily correspond to physical buffers in the origin node of link i . Their introduction enables the differentiation of the traffic, based on the destination and the link the traffic will go through. If, from the topology of the network or from other constraints, it turns out that the traffic of a specific destination never transverses link i , then the corresponding buffer in B_i will be permanently empty.

C. Manufacturing Systems

In manufacturing systems, several different types of parts are fabricated. A part type, in order to be manufactured, needs to be processed by a number of machines in some prespecified order. In flexible manufacturing systems, a machine can process more than one type of part. In this case, the machine has to switch from one part type to another at certain times since the different parts are not processed simultaneously. The switching of the machine involves a period during which the machine idles. The switching should be done rarely enough such that the fraction of time that the machine is utilized is higher than the loading.

Perkins and Kumar [10] have obtained a distributed scheduling policy that stabilizes any acyclic network of manufacturing machines. When cycles are formed by the routes of the parts in the manufacturing system, then the behavior of the system is more difficult to characterize. The intrinsic complexity of the problem was demonstrated by a counterintuitive example of instability in a simple manufacturing system in [6]. The queueing network considered here is readily interpreted into a manufacturing system model. The servers correspond to the machines, the arrival streams to the different part types, and the buffers of each machine store the part types at intermediate processing stages. If a part type needs to be processed by a machine more than once in its manufacturing cycle, then it

waits in a different buffer each time. When the part types have unique routes through the manufacturing system, then the sets \mathcal{R}_j have a unique element, the next buffer that the parts will join when they leave from j . Kumar and Seidman in [6] propose a scheduling policy with a supervisor mechanism that stabilizes the manufacturing system as long as the load of each machine is less than one. The supervisory mechanism, though, relies on knowledge of the arrival rates. The ABP policy stabilizes the system without the arrival rate information. The model considered here extends the one considered in [10], [6] to include cases where a part type has the option to follow different routes at certain points of its processing, in which case routing is performed.

III. NECESSARY AND SUFFICIENT STABILIZABILITY CONDITIONS

We are interested in the average rate with which work can be served. We consider that certain throughput rates can be achieved if there exists a policy under which the network is stable when the long-run average arrival rates are equal to the desired throughputs. Denote by $X_j(t)$ the amount of work in buffer j at the time t .

Definition. The system is stable under a policy π if

$$\limsup_{t \rightarrow \infty} \sum_{j \in \mathcal{B}_i} X_j(t) < D$$

where D may depend only on the arrival rates $\mathbf{a} = (a_1, \dots, a_N)$, the burstiness coefficients $\mathbf{b} = (b_1, \dots, b_N)$, and the service rates $\mu = (\mu_{ij} : i = 1, \dots, N, j \in \mathcal{B}_i)$, but not on the initial condition.

The following condition on the arrival and service rates is necessary and sufficient for the existence of a policy under which the network is stable.

C1. There exist nonnegative numbers f_{jk} , $j = 1, \dots, B$, $k \in \mathcal{R}_j$ that satisfy the flow conservation equations

$$a_i + \sum_{k \in \mathcal{R}_i} f_{ki} = \sum_{k \in \mathcal{R}_i} f_{jk}, \quad i = 1, \dots, B. \quad (3.1)$$

Also, there exist nonnegative numbers u'_j , $i = 1, \dots, N$, $j = 1, \dots, B$ such that

$$\sum_{j \in \mathcal{B}_i} u'_j < 1, \quad i = 1, \dots, N, \quad (3.2a)$$

$$\sum_{i \in \mathcal{R}_j} f_{ji} \leq \sum_{i \in \mathcal{B}_i} u'_i \mu_{ij}, \quad j = 1, \dots, B. \quad (3.2b)$$

More specifically, it is shown that C1 is sufficient for stability if the burstiness condition (2.1) holds, while C1 is necessary for stability when the arrival process satisfies a "bounded idleness" condition, to be specified later. No existence of the arrival rate is needed. An intuitive justification of C1 follows. The number f_{jk} in the condition C1 represents the long-run average rate with which work is transferred from buffer j to buffer k . Hence, (3.1) is indeed a flow conservation equation. A collection of nonnegative numbers f_{jk} , $j = 1, \dots, B$, $k \in \mathcal{R}_j$ which satisfy (3.1) will be referred

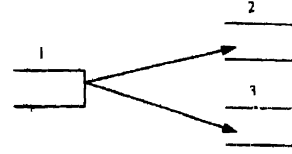


Fig. 3. A network with routing control.

to as flow in the following. The number u'_j represents the fraction of time that server i spends serving buffer j . Inequality (3.2b) expresses the fact that the rate with which work leaves buffer j [the sum on the left side of (3.2b)] cannot be greater than the server capacity [the sum on the right of (3.2b)] that is allocated to that buffer.

In the case of a virtual circuit network, condition C1 is equivalent to the condition that, for every link, the sum of the average rates of all virtual circuits that go through the link is strictly less than its transmission capacity. This can be easily seen by observing the following. Since there is no routing involved, \mathcal{R}_j contains a single element for all j 's and f_{jk} is equal to the arrival rate of the virtual circuit that buffer j corresponds to (recall that in the VC network case, each buffer corresponds to one virtual circuit). Each buffer can be served by one server (link) only; therefore, the sums at the second inequality in (3.2) are reduced to single terms; these inequalities are equivalent to the fact that the traffic load of every link is less than one. The proof of the necessity of C1 for stabilizability follows, while the sufficiency will be proved in Section IV where the parametric back-pressure policy, the predecessor of the ABP, will be specified.

A. Necessity

When the long-run average rate with which work goes from buffer j to k as well as the long-run average fraction of time spent by server i serving buffer j exist, the intuitive interpretation of C1 given earlier turns readily into a proof of the necessity of C1. These long-run averages, though, do not exist for every policy, even if the system is stable, as illustrated in the following counterexample; therefore, a different approach should be taken to show the necessity of C1.

Counterexample: Consider the system in Fig. 3 that consists of three buffers. Buffer 1 receives work with constant rate 0.8, and is served by a server of higher service rate; therefore, it leaves the buffer instantaneously. The served work is routed either to buffer 2 or to 3, from where it leaves the system instantaneously. The switchover time is equal to 0. Clearly, as long as the server does not idle, the system will be stable. The routing is specified by the variable $\tau(t)$ which represents the buffer that is fed from buffer 1 at time t . Let

$$\tau(t) = \begin{cases} 2, & 3^{2k} \leq t \leq 3^{2k+1}, \quad k = 0, 1, \dots \\ 3, & 3^{2k+1} \leq t \leq 3^{2k+2}, \quad k = 0, 1, \dots \end{cases}$$

If f_{12}^i is the average flow from buffer 1 to buffer 2 at the i th routing transition, then we can see that f_{12}^i fluctuates between values which are greater than $\frac{8}{15}$ and less than $\frac{1}{15}$; therefore, it cannot converge, and the long-run average flow does not exist. \diamond

Next, we show the necessity of condition C1. The necessity follows when the arrival streams satisfy the following condition.

S1: There exist nonnegative numbers \hat{b}_i , $i = 1, \dots, B$ such that

$$\int_{t_1}^{t_2} a_i(\tau) d\tau \geq a_i(t_2 - t_1) - \hat{b}_i.$$

Condition S1 can be viewed as a constraint on the idling of the arrival streams.

Theorem 3.1: If the system is stable and condition S1 holds, then condition C1 holds as well.

The proof of the theorem will follow after a lemma. The following definition is needed. A collection of nonnegative numbers f_{jk} , $j = 1, \dots, B$, $k \in \mathcal{R}_j$ is called a *superflow* if

$$a_j + \sum_{k \in \mathcal{R}_k} f_{kj} \leq \sum_{k \in \mathcal{R}_j} f_{jk}, \quad j = 1, \dots, B \quad (3.3)$$

and a *strict superflow* if the inequality in (3.3) is strict whenever $\sum_{k \in \mathcal{R}_j} f_{jk} > 0$. The notions of superflow and strict superflow do not correspond to any physical quantities in the system, and they are introduced only to be used in the proofs.

Lemma 3.1: If $\mathbf{f}' = (f'_{jk}, j = 1, \dots, B, k \in \mathcal{R}_j)$ is a superflow, then there exists a flow $\mathbf{f} = (f_{jk}, j = 1, \dots, B, k \in \mathcal{R}_j)$ such that

$$\mathbf{f} \leq \mathbf{f}' \quad (3.4)$$

where the inequality (3.4) holds componentwise. If \mathbf{f}' is a strict superflow, then there exists a flow \mathbf{f} such that (3.4) holds with strict inequality, for the nonzero elements of \mathbf{f} and \mathbf{f}' .

Proof. Consider the transformation T defined by $\mathbf{f}^{i+1} = T(\mathbf{f}^i)$ where

$$f_{jk}^{i+1} = \begin{cases} \frac{a_j + \sum_{l \in \mathcal{R}_l} f_{lj}^i}{\sum_{l \in \mathcal{R}_j} f_{jl}^i} f_{jk}^i & \text{if } \sum_{l \in \mathcal{R}_j} f_{jl}^i > 0 \\ f_{jk}^i & \text{otherwise.} \end{cases}$$

It is claimed that if \mathbf{f}' is a superflow, then \mathbf{f}^{i+1} is a superflow as well and

$$\mathbf{f}^{i+1} \leq \mathbf{f}' \quad (3.5)$$

while if \mathbf{f}' is a strict superflow, then \mathbf{f}^{i+1} is a strict superflow as well, and (3.5) holds with strict inequality. Notice that (3.5) follows readily since the fact that \mathbf{f}' is a superflow (strict superflow) implies readily that

$$\frac{a_j + \sum_{l \in \mathcal{R}_l} f_{lj}^i}{\sum_{l \in \mathcal{R}_j} f_{jl}^i}$$

is less (strictly less) than one. From the definition of $T(\cdot)$, we have

$$\sum_{k \in \mathcal{R}_j} f_{jk}^{i+1} = a_j + \sum_{l \in \mathcal{R}_l} f_{jl}^i. \quad (3.5a)$$

From (3.5) and (3.5a), it follows that

$$\sum_{k \in \mathcal{R}_j} f_{jk}^{i+1} \geq a_j + \sum_{l \in \mathcal{R}_l} f_{jl}^{i+1} \quad (3.5b)$$

if \mathbf{f}' is a superflow, while (3.5b) holds with strict inequality if \mathbf{f}' is a strict superflow. Relation (3.5b) shows that \mathbf{f}^{i+1} is a superflow and (3.5b), with strict inequality, that it is a strict superflow.

Obviously, if \mathbf{f}' is a flow, then $\mathbf{f}^{i+1} = \mathbf{f}'$. Consider the sequence of superflows \mathbf{f}^i , $i = 0, 1, \dots$ defined as $\mathbf{f}^0 = \mathbf{f}'$, $\mathbf{f}^{i+1} = T(\mathbf{f}^i)$, $i = 0, 1, \dots$. This sequence is nonincreasing, and as a consequence, it converges to a fixed point of $T(\cdot)$ denoted by \mathbf{f}^∞ . Note, though, that all fixed points of T are flows; hence, \mathbf{f}^∞ is a flow which is less than or equal to \mathbf{f}' in general, and strictly less than \mathbf{f}' if \mathbf{f}' is a strict superflow. \diamond

Remark: If \mathbf{f}' in the above lemma contains no cycles, that is, there is no sequence of buffers l_1, l_2, \dots, l_k such that $l_1 = l_k$, $l_{j+1} \in \mathcal{R}_{l_j}$, $j = 1, \dots, k-1$, $f_{l_j l_{j+1}} > 0$, $j = 1, \dots, k-1$, then the sequence of the superflows \mathbf{f}^i , $i = 0, 1, \dots$ in the proof of Lemma 3.1 will converge after a finite number of steps. If there is a cycle in \mathbf{f}' , then the convergence takes an infinite number of steps.

Now, we can proceed to the proof of the theorem.

Proof of Theorem 3.1. Assume that the system is stable. From the definition of stability, there exists D such that for every $\mathbf{X}(0)$, there exists T , which may depend on $\mathbf{X}(0)$, for which

$$\sum X_j(t) \leq D, \quad t \geq T. \quad (3.6)$$

Assume that initially each buffer has a backlog equal to $2D$, and T is such that (3.6) holds for this initial condition. Let Q_{jk} be the total amount of work that has been transferred from buffer j to buffer $k \in \mathcal{R}_j$ in the time interval $(0, T')$, where $T' > T$ is a time to be selected appropriately, as we will see in the following. Clearly, $X_j(T') \leq D$, $j = 1, \dots, B$ and

$$X_j(T') = 2D + \sum_{k \in \mathcal{R}_k} Q_{kj} - \sum_{k \in \mathcal{R}_j} Q_{jk} + \int_0^{T'} a_j(t) dt \leq D, \quad j = 1, \dots, B. \quad (3.6a)$$

Consider the nonnegative vector $\mathbf{f} = (f_{jk} : j = 1, \dots, B, k \in \mathcal{R}_j)$ where $f_{jk} = Q_{jk}/T'$. If we divide (3.6a) by T' , and using S1, we get

$$\begin{aligned} a_j + \sum_{k \in \mathcal{R}_k} f_{kj} &\leq \sum_{k \in \mathcal{R}_j} f_{jk} + a_j - \frac{1}{T'} \int_0^{T'} a_j(t) dt - \frac{D}{T'}, \quad j = 1, \dots, B \\ &\leq \sum_{k \in \mathcal{R}_j} f_{jk} + \frac{\hat{b}_j}{T'} - \frac{D}{T'}. \end{aligned} \quad (3.7)$$

If we select $D > \max_{j=1, \dots, B} \hat{b}_j$, and condition S1 holds, then (3.7) implies that \mathbf{f} is a strict superflow. Let T'_{jk} be the amount of time that server i serves buffer j and directs the traffic to buffer k during the time period from 0 to T' . Define

$$\hat{w}_j = \frac{\sum_{k \in \mathcal{R}_j} T'_{jk}}{T'}.$$

Since each server may serve at most one buffer at a time, we have

$$\sum_{j \in \mathcal{B}_i} \hat{u}_j^i \leq 1 \quad (3.8)$$

We also have

$$Q_{jk} = \sum_{i \in \mathcal{B}} T'_{jk} \mu_{ij} \quad k \in \mathcal{R}_j,$$

which yields

$$\sum_{k \in \mathcal{R}_j} Q_{jk} = \sum_{k \in \mathcal{R}} \sum_{j \in \mathcal{B}} I'_{jk} \mu_{ij} \quad k \in \mathcal{R}_j$$

and therefore

$$\sum_{k \in \mathcal{R}} f_{jk} = \sum_{i \in \mathcal{B}} u_j^i \mu_{ij} \quad k \in \mathcal{R}_j \quad (3.9)$$

From Lemma 3.1, since f is a strict superflow, there exists a flow f' such that $f'_{jk} < f_{jk}$ if $f_{jk} > 0$. Let

$$\epsilon = \max \left\{ \frac{f'_{jk}}{f_{jk}} : j = 1, \dots, B, k \in \mathcal{R}_j, f_{jk} > 0 \right\}$$

Clearly $\epsilon < 1$. Define

$$u_j^i = \epsilon u_j^i \quad i = 1, \dots, n, j \in \mathcal{B}_i$$

It holds that

$$\sum_{j \in \mathcal{B}} u_j^i < 1$$

By multiplying each part of (3.9) by ϵ and given the fact that $f'_{jk} \leq \epsilon f_{jk}$, it follows that

$$\sum_{k \in \mathcal{R}} f'_{jk} \leq \sum_{i \in \mathcal{B}} u_j^i \mu_{ij} \quad j = 1, 2, \dots, B \quad (3.10)$$

◇

The necessity of strict inequality in relation (3.2a) of condition C1 is due to the definition of stability that we have and more specifically to the requirement that the asymptotic bound of the backlog is independent of the initial conditions. It is possible that bounded backlog is guaranteed without strict inequality in (3.2a). This bound, though, cannot be independent of the initial condition. This is illustrated in the following counterexample. Consider the case of a single-server queue with constant instantaneous arrival rate equal to a and service rate μ . Condition C1, with strict inequality in (3.2a), boils down to the condition $a < \mu$. If equality is allowed in (3.2a), we may have $a = \mu$. In the latter case, and if the instantaneous arrival rate is constant and equal to the service rate μ , then the backlog at all times will be equal to the backlog at the time instant 0. Therefore, the network will be unstable according to the definition of stability considered in the paper, since there is no asymptotic bound in the backlog independent of the initial condition.

IV. THE PARAMETRIC BACK PRESSURE POLICY

In this section we will present and study the parametric back-pressure policy $PBP(\alpha)$. This policy determines whether a server idles or not, and if it does not idle, which queue is served and where the traffic is directed. The frequency at which a server is switched from queue to queue depends on the parameter $\alpha > 1$. We include α as an argument in the name of the policy to emphasize that dependence.

PBP(α) While the server i serves a buffer j and directs the traffic to a buffer k , it is constantly monitoring the quantity

$$A_i(t) = \max_{l \in \mathcal{B}} \left\{ \mu_{il} \max_{n \in \mathcal{R}_l} \{ \lambda_l(t) - \lambda_n(t) \} \right\} \quad (4.1)$$

If

$$\alpha \mu_{ij} (\lambda_k(t) - \lambda_j(t)) > 0$$

and

$$A_i(t) \geq \alpha \mu_{ij} (\lambda_k(t) - \lambda_j(t))$$

then the server is rescheduled to serve the queue $l \in \mathcal{B}$ and the traffic is directed to queue $n \in \mathcal{R}_l$ which together realize the maximum in (4.1), with ties broken arbitrarily. Service is reassumed after the switchover time δ .

If

$$A_i(t) \leq 0$$

then the server idles for a time period δ . If at the end of the idling period it is still $A_i(t) \leq 0$, then the server restarts an idling period of the same duration. Otherwise, server i is rescheduled to serve the queue $l \in \mathcal{B}_i$ and the traffic is directed to queue $n \in \mathcal{R}_l$ which together realize the maximum in (4.1).

Some clarifications on the operation of $PBP(\alpha)$ follow. According to $PBP(\alpha)$, the served traffic of buffer l is routed to buffer n which achieves the maximum in

$$\max_{n \in \mathcal{R}_l} \{ \lambda_l(t) - \lambda_n(t) \} \quad (4.1a)$$

whenever buffer l is served. Server i selects which buffer $l \in \mathcal{B}_i$ to serve based on the terms

$$\mu_{il} \max_{n \in \mathcal{R}_l} \{ \lambda_l(t) - \lambda_n(t) \} \quad l \in \mathcal{B}_i$$

It selects the one which achieves the maximum in (4.1). The policy is illustrated in Fig. 4. In order to avoid server i switching from queue to queue too often, the policy reschedules the server only if the quantity (4.1a) for the queue l under service becomes considerably smaller than $A_i(t)$. How much smaller it should become is determined by α which regulates how often the server switches. This feature is reminiscent of the clear a fraction policy considered in [10]. If the backlog in all the downstream queues is larger than the backlog in the queues of \mathcal{B}_i , that is the quantity $A_i(t)$ is negative, then the server i idles.

Note that the scheduling of server i relies on information about the queue lengths of the buffers in \mathcal{B}_i and in $\mathcal{R}_j, j \in \mathcal{B}_i$. This is local information for server i in several practical

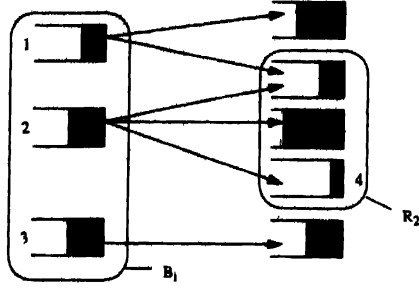


Fig. 4. Server 1 is allocated to one of the queues 1, 2, 3. When the backlogs are as illustrated in the picture, then queue 2 will be selected for service and the traffic will be directed to buffer 4.

systems. For example, in the case of the VC network, the buffers in B_i and in R_j , $j \in B_i$, reside in the origin and destination node of the server/link j ; therefore, policy PBP(α) can be implemented in a distributed fashion.

The stabilizability properties of PBP(α) are stated in the following.

Theorem 4.1: If the arrival and service rates satisfy condition C1, then there exists an $\alpha > 1$ such that the system is stable under PBP(α).

Note that the policy PBP(α) needs only the service rates for its application. Furthermore, the stability holds for any value of the parameter δ in the definition of the policy. The proof of the theorem relies on the following lemma, which is a drift-type condition on the sum of the squared backlogs in the system.

Lemma 4.1: If the arrival and service rates satisfy condition C1, then there exists an α which depends only on the arrival and service rates and the burstiness coefficients such that when policy PBP(α) is employed, the following holds. There exist numbers D_0 , T , and ϵ such that

$$\sum_{j=1}^B X_j^2(t+T) - \sum_{j=1}^B X_j^2(t) < -\epsilon \quad \text{if} \quad \sum_{j=1}^B X_j^2(t) \geq D_0.$$

Proof: The proof of the lemma is lengthy, and relies on some intermediate results. It is given in Appendix A. \diamond

Proof of Theorem 4.1: We show that there exists a \hat{T} , which may depend on $X(0)$, α , μ , and δ , and a D' , which may depend on α , μ , and δ , such that

$$\sup_{t \geq \hat{T}} \sum_{j=1}^B X_j(t) \leq D'. \quad (4.1b)$$

Let T and ϵ be such that

$$\sum_{j=1}^B X_j^2(t+T) - \sum_{j=1}^B X_j^2(t) < -\epsilon \quad \text{if} \quad \sum_{j=1}^B X_j^2(t) \geq D_0. \quad (4.1c)$$

From Lemma 4.1, T and ϵ as above exist. Let \hat{k} be the smallest integer k such that

$$\sum_{j=1}^B X_j^2(kT) \leq D_0.$$

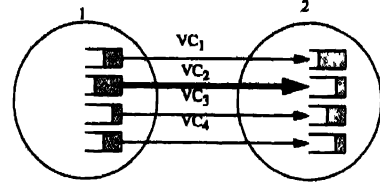


Fig. 5. The link from node 1 to node 2 that carries 4 VC's and the buffers in the origin and destination nodes of the link are illustrated. When the backlogs are as they appear in the picture, the link will transmit VC2.

From (4.1c), clearly such a \hat{k} exists. Define $\hat{T} = \hat{k}T$ and

$$\hat{D} = \left(\sqrt{D_0} + \sum_{j=1}^B (a_j T + b_j) \right)$$

We show by induction that we have

$$\sum_{j=1}^B X_j^2(lT) \leq B\hat{D}, \quad l \geq \hat{k}. \quad (4.1d)$$

The induction step is as follows. If $\sum_{j=1}^B X_j^2(lT) \leq D_0$, then, clearly, $\sum_{j=1}^B X_j^2((l+1)T) \leq B\hat{D}$. If $D_0 \leq \sum_{j=1}^B X_j^2(lT) \leq B\hat{D}$, then $\sum_{j=1}^B X_j^2((l+1)T) \leq \sum_{j=1}^B X_j^2(lT) - \epsilon$, and therefore, $\sum_{j=1}^B X_j^2((l+1)T) \leq \hat{D}$. \diamond

The actions taken by policy PBP(α) in the general network correspond to scheduling the server, routing of the traffic, and idling of the server. If routing is not involved, then the policy is simplified considerably. This is the case in a virtual circuit network where PBP(α) acts as follows. Every link is allocated to the virtual circuits that go through it based on the backlogs of the buffers of these VC's in its origin node and in its destination node. For each VC, the differences of the backlog of the buffer in the origin node of the link minus that of the buffer in the destination node of the link are formed (Fig. 5). If all differences are negative, then the link idles. Otherwise, the quantity

$$A_i(t) = \max_{l \in B_i} \{X(t) - X_l(t)\} \quad (4.1e)$$

is computed where l and \hat{l} are the buffers that correspond to the same virtual circuit in the origin and destination node, respectively, of link i . If

$$A_i(t) > \alpha \mu_{i,j} (X_j(t) - X_{\hat{j}}(t))$$

where j, \hat{j} are the buffers that correspond to the VC currently transmitted, then the server switches to the VC that achieves the maximum in (4.1e).

Note that in the model, it is possible that work goes through the same buffer more than once. It appears that this possibility may lead to instabilities under a distributed policy. The following is an intuitive explanation of why this is not happening with PBP(α). The goal of the policy is to push all the traffic out of the network. It is possible that work may

it a buffer more than once, following a cycle of buffers with small backlogs, instead of following a route out of the network, when the buffers that lead out of the network are congested. Eventually, though, the latter buffers will empty, and the work will find its way out of the network by selecting the route with the low-backlog buffers.

The PBP(α) policy achieves a bounded backlog in the network under the necessary and sufficient stabilizability condition C1, and is indeed distributed since the decisions at each node rely on the one hop away state information. Nevertheless, the arrival and service rates and the topology of the network need to be known in order to select the appropriate α . In the next section, the ABP policy (an adaptive version of the PBP) is obtained, which does not rely on knowledge of the arrival and service rates.

V. THE ADAPTIVE BACK-PRESSURE POLICY

The adaptive back-pressure policy is identical to PBP(α), except that α is not preselected, but is computed at each node based on the local (one hop away) state. Each server i computes a parameter $\alpha_i(t)$ based on the lengths of the queues that it may serve and the queues that it may direct the traffic to, as follows. Consider a function $g: \mathbb{R}^+ \rightarrow \mathbb{R}^+$ which is nonincreasing, strictly greater than 1, and such that

$$\lim_{x \rightarrow \infty} g(x) = 1$$

$$\lim_{x \rightarrow \infty} x(g(x) - 1) = \infty. \quad (5.1)$$

The function $g(x) = 1 + x^{-a}$ satisfies (5.1) for all $0 < a < 1$. Define $\alpha_i(t)$ as

$$\alpha_i(t) = g\left(\max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\}\right).$$

The policy ABP is as follows

ABP: Each server acts as in the PBP(α) policy with the difference that server i uses the locally computed $\alpha_i(t)$ instead of α .

The following holds for ABP.

Theorem 5.1: The network is stable under ABP if condition C1 holds.

The proof of Theorem 5.1 is the same as that of Theorem 4.1, given the drift condition stated in the following lemma.

Lemma 5.1: If the arrival and service rates satisfy condition C1 and the network is operated under ABP, then there exist numbers D_0 , T , and ϵ such that

$$\sum_{j=1}^B X_j^2(t+T) - \sum_{j=1}^B X_j^2(t) < -\epsilon \quad \text{if } \sum_{j=1}^B X_j^2(t) \geq D_0.$$

Proof: The proof is given in Appendix B. \diamond

Note that there is no parameter estimation taking place in the policy. In fact, the average arrival rates might not even exist. The logic of the adaptive policy is that it adjusts α such that it goes closer to 1 with a certain rate as the backlog increases. In

this manner, it is achieved that the servers switch to the queues that have the heavier backlog difference frequently enough.

VI. CONCLUSIONS AND DISCUSSION

In the ABP policy, each server is scheduled based on the lengths of the queues that it serves and of the queues one hop away, which were assumed to be instantaneously available to the server in our study. This is not always the case, though. For example, in the virtual circuit network, the link scheduling was based on the queue lengths at the origin and the destination nodes of the link. Clearly, the queue lengths in the destination node could not be made available at the origin node without a delay greater than or equal to the propagation delay of the link. In high-speed networks, the link propagation time is enough for the state of the queues to change considerably. Therefore, there will be a discrepancy between the actual queue lengths and those available to the server. Nevertheless, as long as the difference between the actual queue lengths and those available to the server is bounded independently of the queue length values, then the results obtained in the paper remain unaffected. That is, the ABP policy stabilizes the network with delayed information about the queue lengths as well.

According to the ABP policy, server i idles if the quantity $A_i(t)$ becomes less than or equal to 0 or, in other words, if for every queue $l \in \mathcal{B}_i$ the backlog of all queues in \mathcal{R}_l is greater than or equal to that of queue l . The results remain unaffected when $A_i(t)$ is compared with an arbitrary negative number instead of 0 in the definition of ABP. That is, the server idles only if the backlog in the downstream queues becomes greater than that of the upstream queues by a certain amount.

The delay through the network is an issue left open for further research. Bounded backlogs imply bounded delay; therefore, under the ABP policy, the delay will be finite when the stability condition C1 holds. There are several questions which are left open, though. How does the delay vary when the function g that estimates the parameter α in each node changes? Which is a good choice of g as far as the delay is concerned? As we said earlier, the backlog in the network remains bounded even if the server i idles whenever $A_i(t) < h$ for an arbitrary h , and not necessarily for $h = 0$. What is a good choice of h for small delays? Also, what will be the effect of the delayed information about the state of the one hop away queues on the delays through the network, and how will the latter be affected from the propagation delay? The investigation of these questions might lead to refinements of ABP with improved performance with respect to delay.

APPENDIX A

Lemma 4.1 is proved here. Two lemmas precede its proof. The first quantifies the property of policy PBP(α) that the switching of the server becomes less frequent as the queue lengths increase and more frequent as α approaches 1. Note that the maximum rate with which the difference of two queues may vary (increase or decrease) is less than or equal to

$$B\lambda + \sum_{i=1}^N \sum_{j \in \mathcal{B}_i} \mu_{ij} = M.$$

Lemma A.1: If for a server i we have

$$\max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \geq \frac{4\alpha}{\alpha - 1} MT \quad (\text{A.1})$$

then i will switch at most once in the time interval $(t, t + T)$ and

$$\begin{aligned} & \sum_{j \in \mathcal{B}_i, l \in \mathcal{R}_j} Q_{ji}^i(X_j(t) - X_l(t)) \\ & \geq \frac{1}{\alpha} (T - \delta) \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ & \quad - 4M(T + \delta)^2 \end{aligned} \quad (\text{A.2})$$

where Q_{ji}^i is the amount of work served by server i and transferred from buffer j to buffer l during the time interval $(t, t + T)$.

Proof: At time t , server i serves queue j_1 and directs the traffic to queue l_1 . We distinguish two cases. Assume first that

$$\mu_{ij_1}(X_{j_1}(t) - X_{l_1}(t)) = \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \quad (\text{A.3})$$

Since the maximum rate with which any difference $\mu_{ij_1}(X_{j_1}(t) - X_{l_1}(t))$ may vary is M , we have

$$\begin{aligned} & \alpha \mu_{ij_1}(X_{j_1}(t') - X_{l_1}(t')) \\ & \geq \alpha \left(\max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} - 2MT \right), \\ & \quad t \in (t, t + T). \end{aligned} \quad (\text{A.4})$$

Also, for any pair of queues $j \in \mathcal{B}_i, l \in \mathcal{R}_j$

$$\begin{aligned} & \mu_{ij}(X_j(t') - X_l(t')) \\ & \geq \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} + 2MT, \\ & \quad t' \in (t, t + T). \end{aligned} \quad (\text{A.5})$$

From (A.4), (A.5) we can see that if (A.1) holds, then

$$\alpha \mu_{ij_1}(X_{j_1}(t') - X_{l_1}(t')) \geq \mu_{ij_1}(X_{j_1}(t') - X_{l_1}(t')), \quad j \in \mathcal{B}_i, j \neq j_1, l \in \mathcal{R}_j, t' \in (t, t + T) \quad (\text{A.5a})$$

and the server will not switch in the time interval $(t, t + T)$. Then, clearly

$$\begin{aligned} & \sum_{j \in \mathcal{B}_i, l \in \mathcal{R}_j} Q_{ji}^i(X_j(t) - X_l(t)) \\ & = Q_{j_1 l_1}^i(X_{j_1}(t) - X_{l_1}(t)) \\ & = T \mu_{ij_1}(X_{j_1}(t) - X_{l_1}(t)) \\ & = T \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \end{aligned} \quad (\text{A.6})$$

and (A.2) follows. Note that (A.6) assumes that the server is always busy during $(t, t + T)$, which is a valid assumption based on (A.1).

Assume now that (A.3) does not hold. From the definition of the policy at time t , we have

$$\alpha \mu_{ij_1}(X_{j_1}(t) - X_{l_1}(t)) \geq \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\}. \quad (\text{A.7})$$

If server i will not switch queues during the interval $(t, t + T)$, then from (A.7), relation (A.2) follows easily. If the server will switch, then let t_1 be the first time after t at which the server switches. Let j_2 be the queue that is served after the switching, and l_2 the queue to which the traffic is routed. At time $t_1 + \delta$, we have

$$\begin{aligned} & \mu_{ij_2}(X_{j_2}(t_1 + \delta) - X_{l_2}(t_1 + \delta)) \\ & = \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t_1 + \delta) - X_l(t_1 + \delta)\} \right\} \\ & \geq \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} - 2M(t_1 - t + \delta) \end{aligned} \quad (\text{A.8})$$

and arguing similarly as above

$$\begin{aligned} & \mu_{ij_2}(X_{j_2}(t') - X_{l_2}(t')) \\ & \geq \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} - 2MT \end{aligned} \quad (\text{A.8a})$$

Equation (A.8a) together with (A.5) imply that if (A.1) holds, then the server will not switch again in the time interval $(t_1 + \delta, t + T)$. Let Q_{ji}^i be the amount of work served by server i and transferred from buffer j to buffer l during the time interval (t, t_1) , let \hat{Q}_{ji}^i be the same quantity for the time interval $(t_1 + \delta, t + T)$. Then, clearly, $Q_{ji}^i = \hat{Q}_{ji}^i + \tilde{Q}_{ji}^i$ and

$$\begin{aligned} & \sum_{j \in \mathcal{B}_i, l \in \mathcal{R}_j} Q_{ji}^i(X_j(t) - X_l(t)) \\ & = (t_1 - t) \mu_{ij_1}(X_{j_1}(t) - X_{l_1}(t)) \\ & \geq \frac{1}{\alpha} (t_1 - t) \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \end{aligned} \quad (\text{A.9})$$

while

$$\begin{aligned} & \sum_{j \in \mathcal{B}_i, l \in \mathcal{R}_j} \tilde{Q}_{ji}^i(X_j(t) - X_l(t)) \\ & = Q_{j_2 l_2}^i(X_{j_2}(t) - X_{l_2}(t)) \\ & = (T + t - t_1 - \delta) \mu_{ij_2}(X_{j_2}(t) - X_{l_2}(t)). \end{aligned} \quad (\text{A.10})$$

Notice that

$$\mu_{ij_2}(X_{j_2}(t) - X_{l_2}(t)) \geq \mu_{ij_2}(X_{j_2}(t_1 + \delta) - X_{l_2}(t_1 + \delta)) - 4M(t_1 - t + \delta) \quad (\text{A.11})$$

From (A.8), (A.10), and (A.11), we get

$$\begin{aligned} & \sum_{j \in \mathcal{B}_i, l \in \mathcal{R}_j} \tilde{Q}_{ji}^i(X_j(t) - X_l(t)) \\ & \geq (T + t - t_1 - \delta) \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ & \quad - 4TM(t_1 - t + \delta). \end{aligned} \quad (\text{A.12})$$

from (A 9) and (A 12), we have

$$\begin{aligned}
 & \sum_{j \in B, l \in \mathcal{R}_j} Q'_{jl}(X_j(t) - X_l(t)) \\
 & \geq \frac{1}{\alpha}(t_1 - t) \max_{j \in B} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\
 & \quad + (T + t - t_1 - \delta) \max_{j \in B} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\
 & \quad - 4TM(t_1 - t + \delta) \\
 & \geq \frac{1}{\alpha}(I - \delta) \max_{j \in B} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\
 & \quad - 4M(I + \delta)^2 \quad (\text{A 12a})
 \end{aligned}$$

In the following we show that one of the differences $X_j(t) - X_l(t)$, $j = 1, \dots, B$, $l \in \mathcal{R}_j$, is on the order of $\sqrt{\sum_{j=1}^B X_j^2(t)}$. A result similar to the following lemma has been shown in [11]. The lemma is included here for completeness. Notice that it holds independently of the policy and is a characteristic of the network.

Lemma A 2 There is a constant $\epsilon > 0$ that depends only on the topology of the network so that

$$\max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \geq \epsilon \sqrt{\sum_{j=1}^B X_j^2(t)}$$

Proof Consider the queue j_0 with the maximum length. Clearly

$$X_{j_0}(t) \geq \sqrt{\sum_{j=1}^B X_j^2(t)/B}$$

There exists a sequence of queues through which the work can be routed out of the network, that is, there exist a sequence i_1, \dots, i_n such that $n \leq B$, $i_1 = j_0$, $0 \in \mathcal{R}_{i_1}$, $i_{k+1} \in \mathcal{R}_{i_k}$, $k = 1, \dots, n-1$. Then we have

$$\begin{aligned}
 & \sum_{k=1}^{n-1} (X_{i_k}(t) - X_{i_{k+1}}(t)) = X_{i_1}(t) - X_{i_n}(t) \\
 & \Rightarrow \max_{k=1, \dots, n-1} (X_{i_k} - X_{i_{k+1}}) \geq \frac{X_{i_1}(t)}{B} \geq \sqrt{\frac{\sum_{j=1}^B X_j^2(t)}{B^3}}
 \end{aligned}$$

Proof of Lemma 4.1 With simple calculations we get for every l

$$\begin{aligned}
 & \sum_{j=1}^B X_j^2(t+I) - \sum_{j=1}^B X_j^2(t) \\
 & = 2 \sum_{j=1}^B (X_j(t+I) - X_j(t)) X_j(t) \\
 & \quad + \sum_{j=1}^B (X_j(t+I) - X_j(t))^2 \quad (\text{A 13})
 \end{aligned}$$

Notice that

$$(X_j(t+I) - X_j(t))^2 \leq M^2 I^2 \quad j = 1, \dots, B$$

where M is the maximum rate with which any X_j vary, and it has been defined before Lemma A 1

$$\sum_{j=1}^B (X_j(t+I) - X_j(t))^2 \leq BM^2 I^2$$

In the following we proceed to bound the first term on the right hand side of (A 13). Let Q'_{jl} be the amount of work served by server j and transferred from buffer j to l during the time interval $(t, t+I)$. Let Q_j be the amount of work transferred to buffer j from outside, and Q_{j0} the amount of work served by server j and transferred from buffer j to outside during $(t, t+I)$. Clearly,

$$X_j(t+I) - X_j(t) + Q_j + \sum_{l \in \mathcal{R}_j} \sum_{i=1}^N Q'_{il} - \sum_{l \in \mathcal{P}_j} \sum_{i=1}^N Q_{il}$$

Hence for the first term on the right side of (A 13) we have

$$\begin{aligned}
 & \sum_{j=1}^B (X_j(t+I) - X_j(t)) X_j(t) \\
 & = \sum_{j=1}^B Q_j X_j(t) + \sum_{j=1}^B \sum_{l \in \mathcal{R}_j} \sum_{i=1}^N Q_{il} X_j(t) \\
 & \quad - \sum_{j=1}^B \sum_{l \in \mathcal{P}_j} \sum_{i=1}^N Q'_{il} X_j(t) \\
 & = \sum_{j=1}^B Q_j X_j(t) + \sum_{j=1}^B \sum_{l \in \mathcal{R}_j} \sum_{i=1}^N Q'_{jl} (X_j(t) - X_l(t)) \quad (\text{A 15})
 \end{aligned}$$

From the burstiness constraints on the arrival streams we have

$$Q_j \leq a_j I + b_j \quad j = 1, \dots, B$$

which together with condition C1 gives

$$\begin{aligned}
 & \sum_{j=1}^B Q_j X_j(t) \leq I \sum_{j=1}^B a_j X_j(t) + \sum_{j=1}^B b_j X_j(t) \\
 & = I \sum_{j=1}^B \left(\sum_{l \in \mathcal{R}_j} f_{jl} + \sum_{l \in \mathcal{P}_j} f_{lj} \right) X_j(t) \\
 & \quad + \sum_{j=1}^B b_j X_j(t) \\
 & = I \sum_{j=1}^B \sum_{l \in \mathcal{R}_j} f_{jl} (X_j(t) - X_l(t)) + \sum_{j=1}^B b_j X_j(t) \\
 & \leq I \sum_{j=1}^B \sum_{l \in \mathcal{R}_j} f_{jl} \max_{k \in \mathcal{R}_j} (X_j(t) - X_k(t)) \\
 & \quad + \sum_{j=1}^B b_j X_j(t) \quad (\text{A 16})
 \end{aligned}$$

In the following, we upper bound the second term on the right side of (A 15). Note that

$$Q'_{il} (X_j(t) - X_l(t)) \geq -M^2 I^2 \quad i = 1, \dots, N, \quad j, l = 1, \dots, B \quad (\text{A 17})$$

This is so because, first

$$0 \leq Q_{jl}^i \leq MT, \quad i = 1, \dots, N, \quad j, l = 1, \dots, B$$

and, second, if $X_j(t) - X_l(t) < -MT$, then $X_j(t') - X_l(t') \leq 0$ for all $t' \in (t, t+T)$, and Q_{jl}^i will be equal to 0 since no server will serve queue j in the time interval $(t, t+T)$. By interchanging the order of the summation, we have

$$\begin{aligned} \sum_{j=1}^B \sum_{l \in \mathcal{R}_j} \sum_{i=1}^N Q_{jl}^i (X_l(t) - X_j(t)) \\ = - \sum_{i=1}^N \sum_{j=1}^B \sum_{l \in \mathcal{R}_j} Q_{jl}^i (X_j(t) - X_l(t)) \quad (\text{A.18}) \end{aligned}$$

Define $\hat{D} = \sum_{j=1}^B X_j^2(t)$. From Lemma A.2, we get that for some i , say $i = i_0$, we have

$$\begin{aligned} \max_{j \in \mathcal{B}_{i_0}} \left\{ \mu_{i_0 j} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ = \max_{i=1, \dots, N} \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ \geq \min_{j \in \mathcal{B}, i=1, \dots, N} \{ \mu_{ij} \} \max_{j \in \mathcal{B}, l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \geq \epsilon \sqrt{D}. \end{aligned}$$

Hence, for \hat{D} large enough, inequality (A.1) is satisfied for at least one server. Let \mathcal{F} be the set of servers for which (A.1) holds. Replacing in (A.18) from (A.2) and (A.17), we get

$$\begin{aligned} \sum_{j=1}^B \sum_{l \in \mathcal{R}_j} \sum_{i=1}^N Q_{jl}^i (X_l(t) - X_j(t)) \\ \leq - \sum_{i \in \mathcal{F}} \frac{1}{\alpha} (T - \delta) \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ + 4M(T + \delta)^2 + NB^2 M^2 T^2 \\ \leq - \sum_{i \in \mathcal{F}} \left(1 - \sum_{j \in \mathcal{B}_i} u_j^i \right) \frac{1}{\alpha} T \max_{j \in \mathcal{B}_i} \\ \cdot \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ + 4M(T + \delta)^2 + NB^2 M^2 T^2 \\ - \sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u_j^i \right) T \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ + \sum_{i \in \mathcal{F}} \frac{\delta}{\alpha} \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ - \sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u_j^i \right) T \left(\frac{1}{\alpha} - 1 \right) \max_{j \in \mathcal{B}_i} \\ \cdot \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \quad (\text{A.19}) \end{aligned}$$

Define

$$u = \min_{i=1, \dots, N} \left\{ 1 - \sum_{j \in \mathcal{B}_i} u_j^i \right\}, \quad m = \min_{i=1, \dots, N} \min_{j \in \mathcal{B}_i} \mu_{ij}. \quad (\text{A.19a})$$

Then, we can easily check that

$$\begin{aligned} \sum_{i \in \mathcal{F}} \left(1 - \sum_{j \in \mathcal{B}_i} u_j^i \right) \frac{1}{\alpha} T \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ \geq u \frac{1}{\alpha} T m \max_{j=1, \dots, B, l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\}. \quad (\text{A.20}) \end{aligned}$$

Clearly

$$\sqrt{\hat{D}} \geq \max_{j=1, \dots, B, l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \quad (\text{A.21})$$

Let $\mu = \max_{i=1, \dots, N} \max_{j \in \mathcal{B}_i} \mu_{ij}$. Then, from (A.21)

$$\max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \leq \mu \sqrt{\hat{D}}, \quad i = 1, \dots, N$$

and

$$\sum_{i \in \mathcal{F}} \frac{\delta}{\alpha} \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \leq N \frac{\delta}{\alpha} \mu \sqrt{\hat{D}} \quad (\text{A.22})$$

From (A.21), and since $1/\alpha < 1$, we have

$$\begin{aligned} \sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u_j^i \right) T \left(\frac{1}{\alpha} - 1 \right) \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ \geq \sum_{i \in \mathcal{F}} T \left(\frac{1}{\alpha} - 1 \right) 2\mu \sqrt{\hat{D}} \\ \geq 2NT \left(\frac{1}{\alpha} - 1 \right) \mu \sqrt{\hat{D}} \quad (\text{A.23}) \end{aligned}$$

From (A.19), (A.20), (A.22), and (A.23), we have

$$\begin{aligned} \sum_{j=1}^B \sum_{l \in \mathcal{R}_j} \sum_{i=1}^N Q_{jl}^i (X_l(t) - X_j(t)) \\ \leq -u \frac{1}{\alpha} T m \epsilon \sqrt{\hat{D}} + 4M(T + \delta)^2 + NB^2 M^2 T^2 \\ + N \frac{\delta}{\alpha} \mu \sqrt{\hat{D}} - 2NT \left(\frac{1}{\alpha} - 1 \right) \mu \sqrt{\hat{D}} \\ - \sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u_j^i \right) T \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \quad (\text{A.24}) \end{aligned}$$

Let $\hat{\mathcal{F}}$ be the set of all queues j for which

$$\min_{l \in \mathcal{R}_j} (X_j(t) - X_l(t)) > \frac{4\alpha}{(\alpha - 1)} MT \quad (\text{A.24a})$$

Then, from (A.16), we have

$$\begin{aligned} \sum_{j=1}^B Q_j X_j(t) \\ \leq \sum_{j=1}^B X_j(t) b_j + T \sum_{j \in \hat{\mathcal{F}}} \max_{k \in \mathcal{R}_j} \{X_j(t) - X_k(t)\} \sum_{l \in \mathcal{R}_j} f_{jl} \\ + B \sum_{i=1}^B a_i \frac{4\alpha}{\alpha - 1} MT. \quad (\text{A.25}) \end{aligned}$$

ote that from (A.24a), if $j \in \hat{\mathcal{F}}$, then all i 's such that $j \in \mathcal{B}_i$ long to \mathcal{F} . Based on this fact, in the following, we verify that

$$\sum_{i \in \mathcal{F}} \max_{k \in \mathcal{R}_i} \{X_j(t) - X_k(t)\} \sum_{l \in \mathcal{R}_i} f_{jl} \sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u'_j \right) T \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_i} \{X_j(t) - X_l(t)\} \right\}. \quad (\text{A.26})$$

Note first that

$$\sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u'_j \right) T \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_i} \{X_j(t) - X_l(t)\} \right\} \geq T \sum_{i \in \mathcal{F}} \sum_{j \in \mathcal{B}_i} u'_j \mu_{ij} \max_{l \in \mathcal{R}_i} \{X_j(t) - X_l(t)\} \quad (\text{A.27})$$

and from the fact that if $j \in \hat{\mathcal{F}}$, then all i 's such that $j \in \mathcal{B}_i$ belong to \mathcal{F} , we have

$$I \sum_{i \in \mathcal{F}} \sum_{j \in \mathcal{B}_i} u'_j \mu_{ij} \max_{l \in \mathcal{R}_i} \{X_j(t) - X_l(t)\} \geq T \sum_{i \in \mathcal{F}} \sum_{j \in \mathcal{B}_i} u'_j \mu_{ij} \max_{l \in \mathcal{R}_i} \{X_j(t) - X_l(t)\}. \quad (\text{A.28})$$

From (A.27), (A.28), and (3.2), the relation (A.26) follows. By adding the inequalities (A.24)–(A.26) and (A.15), and after the simplifications, we get

$$\begin{aligned} & \sum_{j=1}^B (X_j(t+T) - X_j(t)) X_j(t) \\ & \leq -u \frac{1}{\alpha} T \mu \sqrt{D} + 4M(I + \delta)^2 + NB^2 M^2 T^2 \\ & \quad + N \frac{\delta}{\alpha} \mu \sqrt{D} - 2NT \left(\frac{1}{\alpha} - 1 \right) \mu \sqrt{D} \\ & \quad + \sqrt{D} \sum_{j=1}^B b_j + B \sum_{i=1}^B a_i \frac{\alpha \alpha}{\alpha - 1} MT \end{aligned} \quad (\text{A.28a})$$

By replacing in (A.13) from (A.14) and (A.28a), we get

$$\begin{aligned} & \sum_{j=1}^B X_j^2(t+T) - \sum_{j=1}^B X_j^2(t) \\ & \leq c_1 T^2 - c_2 \frac{1}{\alpha} T \sqrt{\hat{D}} + c_3 \sqrt{D} \\ & \quad + c_4 T \sqrt{\hat{D}} \left(1 - \frac{1}{\alpha} \right) \\ & \quad + c_5 \frac{4\alpha}{\alpha - 1} T \end{aligned} \quad (\text{A.29})$$

where c_1, \dots, c_5 are positive constants which depend only on the system topology and the arrival and service parameters. Select $\alpha > 1$ such that $c_4 - (c_2 + c_4/\alpha) = -\zeta < 0$, and let $T = (\hat{D})^{1/4}$. Then (A.29) becomes

$$\begin{aligned} & \sum_{j=1}^B X_j^2(t+T) - \sum_{j=1}^B X_j^2(t) \\ & \leq c_1 (\hat{D})^{1/2} - \zeta (\hat{D})^{3/4} + c_3 \sqrt{D} \\ & \quad + c_5 \frac{4\alpha}{\alpha - 1} (\hat{D})^{1/4} \end{aligned}$$

It is clear that if \hat{D} is large enough, the right side of (A.29) becomes strictly negative and the lemma follows.

APPENDIX B

The proof of Lemma 5.1 is given in this Appendix after some intermediate results. Define

$$\alpha_i^T = q \left(\max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_i} \{X_j(t) - X_l(t)\} \right\} + 2MI \right)$$

Lemma B.1 If for a server i , we have

$$\max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_i} \{X_j(t) - X_l(t)\} \right\} > \frac{4\alpha_i^T}{(\alpha_i^T - 1)} MT \quad (\text{B.1})$$

then i will switch at most once in the time interval $(t, t+T)$ and

$$\begin{aligned} & \sum_{j \in \mathcal{B}} \sum_{l \in \mathcal{R}_i} Q'_{jl} (X_j(t) - X_l(t)) \\ & \geq \frac{1}{\alpha_i(t)} (T - \delta) \max_{j \in \mathcal{B}} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_i} \{X_j(t) - X_l(t)\} \right\} \\ & \quad - 4M(T + \delta)^2 \end{aligned} \quad (\text{B.2})$$

where Q'_{jl} is the amount of work served by server i and transferred from buffer j to buffer l during the time interval $(t, t+T)$.

Proof The proof follows the same steps as with Lemma A.1, except for the following differences. Inequality (A.4) holds with α replaced by $\alpha_i(t)$ on the left side and with α_i^T on the right side. Inequalities (A.5a), (A.7), (A.9), and (A.12a) hold with α replaced by $\alpha_i(t)$. \diamond

Proof of Lemma 5.1 From (A.13)–(A.15), we get

$$\begin{aligned} & \sum_{j=1}^B X_j^2(t+T) - \sum_{j=1}^B X_j^2(t) \\ & = \sum_{j=1}^B Q_j X_j(t) \\ & \quad + \sum_{j=1}^B \sum_{l \in \mathcal{R}_i} \sum_{i=1}^N Q'_{jl} (X_l(t) - X_j(t)) \\ & \quad + BM^2 T^2. \end{aligned} \quad (\text{B.3})$$

Also

$$\begin{aligned} & \sum_{j=1}^B Q_j X_j(t) \leq T \sum_{j=1}^B \sum_{l \in \mathcal{R}_i} f_{jl} \max_{k \in \mathcal{R}_i} (X_j(t) - X_k(t)) \\ & \quad + \sum_{j=1}^B X_j(t) b_j. \end{aligned} \quad (\text{B.4})$$

In the following, we upper bound the second term on the right side of (B.3). Note first that

$$Q'_{jl}(X_j(t) - X_l(t)) \geq -M^2 T^2, \quad i = 1, \dots, N, \\ j, l = 1, \dots, B. \quad (\text{B.5})$$

By interchanging the order of the summation, we have

$$\sum_{j=1}^B \sum_{l \in \mathcal{R}_j} \sum_{i=1}^N Q'_{il}(X_i(t) - X_j(t)) \\ = - \sum_{i=1}^N \sum_{j=1}^B \sum_{l \in \mathcal{R}_j} Q'_{il}(X_j(t) - X_l(t)). \quad (\text{B.6})$$

Let \mathcal{F} be the set of servers for which (B.1) holds. Replacing in (B.6) from (A.2) and (B.5), we get

$$\sum_{j=1}^B \sum_{l \in \mathcal{R}_j} \sum_{i=1}^N Q'_{il}(X_i(t) - X_j(t)) \\ \leq - \sum_{i \in \mathcal{F}} \frac{1}{\alpha_i(t)} (T - \delta) \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ + 4M(T + \delta)^2 + NB^2 M^2 T^2 \\ \leq - \sum_{i \in \mathcal{F}} \left(1 - \sum_{j \in \mathcal{B}_i} u'_j \right) \frac{1}{\alpha_i(t)} T \max_{j \in \mathcal{B}_i} \\ \cdot \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ + 4M(T + \delta)^2 + NB^2 M^2 T^2 \\ - \sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u'_j \right) T \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ + \sum_{i \in \mathcal{F}} \frac{\delta}{\alpha_i(t)} \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ - \sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u'_j \right) T \left(\frac{1}{\alpha_i(t)} - 1 \right) \max_{j \in \mathcal{B}_i} \\ \cdot \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \quad (\text{B.7})$$

To simplify the right side of inequality (B.7), we upper bound the terms appearing in it in the following. Based on the definition of u and m in (A.19a), the nonincreasingness of

$g(\cdot)$, and the definition of $\alpha_i(t)$, we have

$$\sum_{i \in \mathcal{F}} \left(1 - \sum_{j \in \mathcal{B}_i} u'_j \right) \frac{1}{\alpha_i(t)} T \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ \geq u \frac{1}{g(m \max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\})} T m \\ \times \max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\}. \quad (\text{B.8})$$

Since $\alpha_i(t) > 1$, $i = 1, \dots, N$, $t > 0$, we have

$$\sum_{i \in \mathcal{F}} \frac{\delta}{\alpha_i(t)} \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ \leq N \delta \mu \max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \quad (\text{B.9})$$

Similarly, since $1/\alpha_i(t) < 1$, $i = 1, \dots, N$, $t > 0$, $\sum_{j \in \mathcal{B}_i} u'_j \leq 1$, and for μ as defined in (A.19a), we have (B.10), as found at the bottom of the page. From the definition of \mathcal{F} and (B.1), (B.10) becomes

$$\sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u'_j \right) T \left(\frac{1}{\alpha_i(t)} - 1 \right) \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ \geq NT \left(\frac{1}{g((2\alpha_i^T/\alpha_i^I - 1)MT)} - 1 \right) \mu \\ \cdot \max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \\ \geq NT \left(\frac{1}{g(2MT)} - 1 \right) \mu \max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \quad (\text{B.11})$$

From (B.7)–(B.9) and (B.11), we have

$$\sum_{j=1}^B \sum_{l \in \mathcal{R}_j} \sum_{i=1}^N Q'_{il}(X_i(t) - X_j(t)) \\ \leq -u \frac{1}{g(m \max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\})} T m \\ \times \max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} + 4M(T + \delta)^2 \\ + NB^2 M^2 T^2 + N \delta \mu \max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \\ - NT \left(\frac{1}{g(2MT)} - 1 \right) \mu \max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \\ - \sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u'_j \right) T \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\}. \quad (\text{B.12})$$

$$\sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u'_j \right) T \left(\frac{1}{\alpha_i(t)} - 1 \right) \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \right\} \\ \geq NT \left(\frac{1}{\max_{i \in \mathcal{F}} \alpha_i(t)} - 1 \right) 2\mu \max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \\ \geq NT \left(\frac{1}{g(\min_{i \in \mathcal{F}} \max_{j \in \mathcal{B}_i} \{ \mu_{ij} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\} \})} - 1 \right) \mu \max_{j=1, \dots, B} \max_{l \in \mathcal{R}_j} \{X_j(t) - X_l(t)\}. \quad (\text{B.10})$$

let $\hat{\mathcal{F}}$ be the set of all queues j for which

$$\min_{l \in \mathcal{R}_j} (X_j(t) - X_l(t)) > \frac{2\alpha_l^I}{(\alpha_l^T - 1)} MT \quad (\text{B } 13)$$

Then from (B 4), we have

$$\begin{aligned} \sum_{j=1}^B Q_j X_j(t) &\leq \sum_{j=1}^B \lambda_j(t) b_j \\ &\quad + T \sum_{j \in \mathcal{F}} \max_{k \in \mathcal{R}} \{X_j(t) - X_k(t)\} \sum_{l \in \mathcal{R}} f_{jl} \\ &\quad + B \sum_{i=1}^B a_i \frac{2\alpha_i^I}{\alpha_i^T - 1} MT \end{aligned} \quad (\text{B } 14)$$

Notice that

$$\begin{aligned} B \sum_{i=1}^B a_i \frac{2\alpha_i^I}{\alpha_i^T - 1} MT &< 2BMT \sum_{i=1}^B a_i + 2BMT \sum_{i=1}^B a_i \frac{1}{\min_{i=1}^B \alpha_i^T - 1} \\ &\leq 2BMT \sum_{i=1}^B a_i + 2BMT \sum_{i=1}^B a_i \frac{1}{q(\mu \max_{j=1}^B \frac{1}{B \sum_{l \in \mathcal{R}} \{X_j(t) - X_l(t)\}} - 1)} \end{aligned} \quad (\text{B } 15)$$

Defining $D = \sum_{j=1}^B X_j^2(t)$ we have

$$\sum_{j=1}^B \lambda_j(t) b_j \leq \sum_{j=1}^B B_j \sqrt{D} \quad (\text{B } 16)$$

From (B 15) and (B 16) relation (B 14) becomes

$$\begin{aligned} \sum_{j=1}^B Q_j X_j(t) &\leq \sum_{j=1}^B b_j \sqrt{D} + 2BMT \sum_{i=1}^B a_i + 2BMT \sum_{i=1}^B a_i \frac{1}{q(\mu \max_{j=1}^B \frac{1}{B \sum_{l \in \mathcal{R}} \{X_j(t) - X_l(t)\}} + 2MT) - 1} \\ &\quad + T \sum_{j \in \mathcal{F}} \max_{k \in \mathcal{R}} \{X_j(t) - X_k(t)\} \sum_{l \in \mathcal{R}} f_{jl} \end{aligned} \quad (\text{B } 17)$$

Note that from (B 13), if $j \in \mathcal{F}$ then all l 's such that $j \in \mathcal{B}_l$ belong to \mathcal{F} . Based on this fact, and arguing similarly as in the proof of Lemma 4.1 [relations (A 27), (A 28)], it easily follows that

$$\begin{aligned} T \sum_{j \in \mathcal{F}} \max_{k \in \mathcal{R}} \{X_j(t) - X_k(t)\} \sum_{l \in \mathcal{R}} f_{jl} &\leq \sum_{i \in \mathcal{F}} \left(\sum_{j \in \mathcal{B}_i} u_j' \right) T \max_{j \in \mathcal{B}_i} \left\{ \mu_{ij} \max_{l \in \mathcal{R}} \{X_j(t) - X_l(t)\} \right\} \end{aligned} \quad (\text{B } 18)$$

By replacing in (B 3) from (B 12), (B 17) and using (1) and Lemma B 1, we get

$$\begin{aligned} \sum_{j=1}^B X_j^2(t+T) - \sum_{j=1}^B X_j^2(t) &\leq -u \frac{1}{q(\mu \sqrt{D})} T \mu \sqrt{D} \\ &\quad + 1M(I+\delta)^2 + NB^2 \lambda^2 T \\ &\quad + N\delta \mu \sqrt{D} - N T \left(\frac{1}{q(2MT)} - 1 \right) \mu \sqrt{D} \\ &\quad + BM^2 T^2 + \sum_{i=1}^B b_i \sqrt{D} + 2BMT \sum_{i=1}^B a_i \\ &\quad + BMT \sum_{i=1}^B a_i \frac{1}{q(\mu \sqrt{D} + 2MT) - 1} \end{aligned} \quad (\text{B } 19)$$

If we select $I = (D)^{1/4}$ then (B 19) becomes

$$\begin{aligned} \sum_{j=1}^B X_j^2(t+T) - \sum_{j=1}^B X_j^2(t) &\leq \epsilon_1 D^{1/4} + \epsilon_2 \sqrt{D} - \epsilon_3 \frac{1}{q(\epsilon_4 \sqrt{D})} D^{1/4} \sqrt{D} \\ &\quad - \epsilon_5 \left(\frac{1}{q(2MD^{1/2})} - 1 \right) D^{1/4} \sqrt{D} \\ &\quad + \epsilon_6 D^{1/4} \frac{\sqrt{D}}{\sqrt{D}(q(\epsilon_7 \sqrt{D} + \epsilon_8 D^{1/4}) - 1)} \end{aligned} \quad (\text{B } 20)$$

where $\epsilon_1, \dots, \epsilon_8$ are constants that depend only on the system parameters. Since $\lim_{t \rightarrow \infty} q(t) = 1$ there exists $\epsilon_4 > 0$ such that for D sufficiently large,

$$-\epsilon_3 \frac{1}{q(\epsilon_4 \sqrt{D})} - \epsilon_5 \left(\frac{1}{q(2MD^{1/2})} - 1 \right) < -\epsilon_9 < 0 \quad (\text{B } 21)$$

Also there exists a ϵ_{10} such that for D sufficiently large

$$\epsilon_1 D^{1/4} + \epsilon_2 \sqrt{D} \leq \epsilon_{10} \sqrt{D} \quad (\text{B } 22)$$

Hence (B 20) becomes

$$\begin{aligned} \sum_{j=1}^B X_j^2(t+T) - \sum_{j=1}^B X_j^2(t) &\leq \epsilon_{10} \sqrt{D} - \epsilon_9 D^{3/4} + \epsilon_6 D^{1/4} \frac{D^{1/2}}{D^{1/2}(q(\epsilon_7 \sqrt{D} + \epsilon_8 D^{1/4}) - 1)} \\ &= \epsilon_{10} \sqrt{D} + D^{3/4} \left(\epsilon_6 \frac{1}{\sqrt{D}(q(\epsilon_7 \sqrt{D} + \epsilon_8 D^{1/4}) - 1)} - \epsilon_{10} \right) \end{aligned} \quad (\text{B } 23)$$

Since $\lim_{t \rightarrow \infty} t(q(t) - 1) = \infty$ we have that for D large enough, there exists $\epsilon_{11} > 0$ such that

$$\epsilon_6 \frac{1}{\sqrt{D}(q(\epsilon_7 \sqrt{D} + \epsilon_8 D^{1/4}) - 1)} - \epsilon_{10} < -\epsilon_{11} \quad (\text{B } 24)$$

From (B.23) and (B.24), we have for large \hat{D} that

$$\sum_{j=1}^B X_j^2(t) \leq c_{10} \sqrt{\hat{D}} - c_{11} \hat{D}^{3/4}. \quad (\text{B.25})$$

It is clear from (5.1) that if \hat{D} is large enough, then the right side of (B.25) becomes strictly negative, and the lemma follows. \diamond

ACKNOWLEDGMENT

The author would like to thank the reviewers for the thorough reviews that helped improve the final version of the paper.

REFERENCES

- [1] C.-S. Chang, "Stability, queue length, and delay, Part I: Deterministic queueing networks," Tech. Rep. RC 17708, IBM T. J. Watson Res. Cen., Yorktown Heights, NY, 1992.
- [2] R. L. Cruz, "A calculus of network delay, Part I: Network elements in isolation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 114–131, Jan. 1991.
- [3] ———, "A calculus of network delay, Part II: Network analysis," *IEEE Trans. Inform. Theory*, vol. 37, pp. 132–141, Jan. 1991.
- [4] E. L. Hahne, "Round-robin scheduling for max-min fairness in data networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 1024–1039, Sept. 1991.
- [5] M. G. H. Katevenis, "Fast switching and fair control of congested flow in broadband networks," *IEEE J. Select. Areas Commun.*, vol. SAC-5, pp. 1315–1326, Oct. 1987.
- [6] P. R. Kumar and T. L. Seidman, "Distributed instabilities and stabilization methods in distributed real time scheduling of manufacturing systems," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 289–298, 1989.
- [7] S. H. Lu and P. R. Kumar, "Distributed scheduling based on due dates and buffer priorities," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 289–298, 1991.
- [8] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated service networks: The single node case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 344–357, 1993.
- [9] S. S. Panwar, T. K. Philips, and M. S. Chen, "Golden ratio scheduling for flow control with low buffer requirements," *IEEE Trans. Commun.*, vol. 40, pp. 765–772, Apr. 1992.
- [10] J. Perkins and P. R. Kumar, "Stable, distributed, real-time scheduling of flexible manufacturing assembly/disassembly systems," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 139–148, Feb. 1989.
- [11] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling for maximum throughput in multihop radio networks," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1936–1948, Dec. 1992.
- [12] ———, "Throughput properties of a queueing network with distributed dynamic routing and flow control," *Advances Applied Probability*, March, 1994.
- [13] O. Yaron and M. Sidi, "Calculating performance bounds in communication networks," in *IEEE INFOCOM*, 1993, pp. 539–546.
- [14] L. Zhang, "Virtual clock: A new traffic control algorithm for packet switched networks," presented at SIGCOM'90, Philadelphia, PA, 1990.



Leandros Tassiulas (S'89–M'82) was born in 1965 in Katerini, Greece. He received the Diploma in electrical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece in 1987 and the M.S. and Ph.D. degrees, also in electrical engineering, from the University of Maryland, College Park, in 1989 and 1991, respectively.

Since 1991, he has been an Assistant Professor in the Department of Electrical Engineering, Polytechnic University, Brooklyn, NY. His research interests include computer and communication networks with

an emphasis on wireless communications and high-speed networks, the control and optimization of stochastic systems, and parallel and distributed processing.

Dr. Tassiulas coauthored a paper that received the INFOCOM '94 Best Paper Award.

Stability of Queueing Networks and Scheduling Policies

P. R. Kumar, *Fellow, IEEE*, and Sean P. Meyn, *Member, IEEE*

Abstract—Usually, the stability of queueing networks is established by explicitly determining the invariant distribution. Outside of the narrow class of queueing networks possessing a product form solution, however, such explicit solutions are rare, and consequently little is also known concerning stability.

We develop here a programmatic procedure for establishing the stability of queueing networks and scheduling policies. The method uses linear or nonlinear programming to determine what is an appropriate quadratic functional to use as a Lyapunov function. If the underlying system is Markovian, our method establishes not only positive recurrence and the existence of a steady-state probability distribution, but also the geometric convergence of an exponential moment.

We illustrate this method on several example problems. For an example of an open re-entrant line, we show that all stationary nonidling policies are stable for all load factors less than one. This includes the well-known First Come First Serve (FCFS) policy. We determine a subset of the stability region for the Dai-Wang example, for which they have shown that the Brownian approximation does not hold. In another re-entrant line, we show that the Last Buffer First Serve (LBFS) and First Buffer First Serve (FBFS) policies are stable for all load factors less than one. Finally, for the Rybko-Stolyar example, for which a subset of the instability region has been determined by them under a certain buffer priority policy, we determine a subset of the stability region.

I. INTRODUCTION

USUALLY the stability of a queueing network is established by explicitly determining an invariant distribution. Outside of the relatively narrow class of queueing networks admitting a product form solution for the invariant distribution, however, such explicit solutions are rare. Consequently, stability results are also rare.

Here we develop a procedure for establishing the stability of a queueing network operating under a scheduling policy. It is based on just solving a linear program on the coefficients of a quadratic form. Alternatively, a nonlinear program can also be used. The goal is to programmatically construct a quadratic Lyapunov function on buffer levels that has a negative drift, whenever the mean number of parts in the system is large. This allows one to deduce the stability-in-the-mean of a system, even if it is not Markovian. For Markovian systems, such stability is equivalent to the existence of a steady-state

distribution, i.e., positive recurrence. Moreover, for Markovian systems, our method also establishes geometric convergence of an exponential moment.

Such stability results are important for a variety of reasons. First, they are a precursor to more fine grained questions concerning the performance levels of various scheduling policies. Second, several unstable scheduling policies have recently been discovered. Kumar and Seidman [2] and Chase and Ramadge [3] provide examples of deterministic systems which are unstable under all clearing, i.e., exhaustive service, policies. Lu and Kumar [1] provide an example of a re-entrant line with deterministic processing times, and deterministic bursty arrivals, for which a certain buffer priority policy is unstable. Rybko and Stolyar [4] provide an example of a stochastic network which is unstable under a certain buffer priority policy. Recently, Seidman [5] has demonstrated the instability of the well-known First Come First Serve (FCFS) policy, also for a deterministic model. Bramson [6] has recently constructed a stochastic re-entrant line that is also unstable under the FCFS policy. Third, there has been much recent interest in the use of heavy traffic Brownian approximations to construct scheduling policies for queueing networks, see Harrison [7] and Harrison and Wein [8]. Clearly, to establish heavy traffic limit theorems, it is necessary to establish the stability of the queueing networks involved. Dai and Wang [12] have constructed a counterexample where the Brownian approximation does not hold, see also Whitt [13] and Dai and Nguyen [14]. Indeed, heavy traffic limit theorems appear to be only available for systems that are already known to be stable, see Reiman [9], [10] and Peterson [11].

Quadratic Lyapunov functions find widespread use in linear system theory. For stochastic systems, Kingman [15] has used a quadratic Lyapunov function to analyze a random walk on \mathbb{Z}_+^2 . Fayolle [16] has used general quadratic forms to characterize ergodicity of random walks on \mathbb{Z}_+^n . Piecewise linear Lyapunov functions are used in Fayolle *et al.* [17] for establishing the stability of Jackson networks [19]. Meyn and Down [18] use the square of the workload to establish the stability of generalized Jackson networks, where the assumptions on the arrival processes and service times are relaxed. Coffman *et al.* [20] have used linear programming to find both linear and quadratic forms with negative or positive drift and thus the stability or instability of a certain bin packing algorithm. Then linear programs test for a drift of the appropriate sign at all the states on the boundary of a prescribed hypercube, our approach may be less computationally complex. Recently, Bertsimas *et al.* [21] and Kumar and Kumar [22] have used

Manuscript received June 4, 1993; revised March 28, 1994. Recommended by Associate Editor S. Lafontaine. This work was supported in part by NSF Grants ECS 90-25007 and ECS 92-16487 and by the USARO Contract DAAI-03-91-G-0182.

The authors are with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801, USA.

IEEE Log Number 9407565

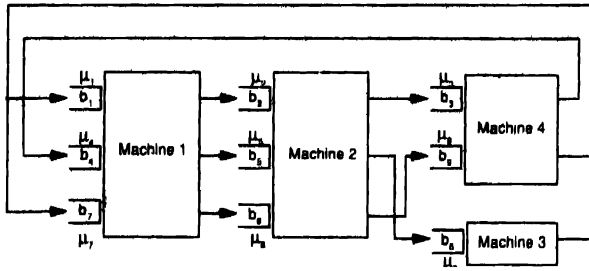


Fig. 1. The basic open re-entrant line.

quadratic forms to obtain performance bounds for queueing networks and scheduling policies, provided the system is stable and has a bounded second moment.

II. THE BASIC OPEN RE-ENTRANT LINE

To expose the idea in its simplest form, we begin with the treatment of an open re-entrant line; see [23].

The network consists of S machines $\{1, 2, \dots, S\}$; see Fig. 1. Parts arrive as a Poisson process of rate λ to buffer b_1 , located at machine $\sigma(1) \in \{1, \dots, S\}$. Upon completing service, they proceed to buffer b_2 located at machine $\sigma(2) \in \{1, \dots, S\}$. Let b_L at machine $\sigma(L)$ be the last buffer visited. The sequence $\{\sigma(1), \dots, \sigma(L)\}$ is the route of the part. Since one may have $\sigma(i) = \sigma(j)$ for some pairs i and j with $i \neq j$, we say that the system is a re-entrant line. Let us suppose that parts in b_i require an exponentially distributed service time, with mean $\frac{1}{\mu_i}$, from machine $\sigma(i)$.

Let $I(i) := \{j \mid \sigma(j) = \sigma(i)\}$, i.e., the set of indexes of buffers which are located at the same machine as b_i . Thus, the buffers with indexes in $I(i)$ are in "contention" for the same machine.

Let us denote

$x_i(t) :=$ Numbers of parts in buffer b_i at time t , including any in service

and

$w_i(t) :=$ 1 if machine $\sigma(i)$ is working on a part in buffer b_i at time t , and zero otherwise.

For simplicity, we suppose that a machine works on only one part at any given time.

The key problem in "scheduling" such queueing networks is to determine which part in which buffer the machine should serve, i.e., which $w_i(t)$ should be one. Clearly, an optimal choice for reasonable criteria will depend on the location, i.e., the buffer, occupied by the part. When the service priority depends on the buffer (i.e., "class" of a part), however, the steady-state distribution, if any, is not known, and, as mentioned earlier, neither is stability. As an example, the well-known FCFS policy can be stable or unstable for particular systems, when the μ_i 's are not the same for all buffers at a machine. Similarly, buffer priority policies can be stable or unstable for particular systems and values of parameters.

III. COPOSITIVE MATRICES

Let Q be a symmetric $L \times L$ matrix that gives rise to a quadratic form which is nonnegative on the positive orthant

in R^L , i.e.,

$$Q = Q^T \text{ and } (y_1, \dots, y_L)Q(y_1, \dots, y_L)^T \geq 0$$

whenever

$$y_i \geq 0 \text{ for all } i. \quad (1)$$

Such matrices are called symmetric copositive matrices. As will be shown in Section V, our methodology will automatically confine itself to the subclass of symmetric strictly copositive matrices. These are symmetric copositive matrices Q for which additionally

$$(y_1, \dots, y_L)Q(y_1, \dots, y_L)^T > 0, \text{ if } y_i \geq 0 \text{ for all } i, \text{ and } y_i \neq 0 \text{ for some } i.$$

Copositive and strictly copositive matrices have been extensively studied; see Cottle *et al.* [24]. They are characterized by the signs of certain determinants (see [24]).

Keller's Theorem: A symmetric matrix is copositive if and only if each principal submatrix, for which the cofactors of the last row are nonnegative, has a nonnegative determinant.

A recent algorithm for testing copositivity can be found in Andersson *et al.* [25]. The determination of copositivity is NP-Complete; see Murty and Kabadi [26].

The procedure we advocate below could be used with any Q satisfying (1). For concreteness, we will confine our attention to the following special types of copositive matrices.

It is easy to see that any symmetric, nonnegative matrix, i.e., one for which $Q^T = Q = [q_{ij}]$, with $q_{ij} \geq 0$ for all i, j , is copositive. Also, any positive semidefinite matrix, i.e., a Q for which, $Q = Q^T$ and $x^T Q x \geq 0$ for all x , is copositive. Moreover, any convex combination (or linear combination with positive weights) of such matrices is also copositive, since the set of symmetric copositive matrices is a convex cone.

IV. THE BASIC CHARACTERIZATION

We shall rescale time so that

$$\lambda + \sum_{i=1}^L \mu_i = 1 \quad (2)$$

and resort to "uniformization," see Lippman [27]. That is, we shall suppose that there is always either a real or a "virtual" part that is being served at every buffer b_i . Let $\{\tau_n\}$, with $\tau_0 = 0$, denote the sequence of all arrival and service times, real or virtual, and let \mathcal{F}_{τ_n} denote the σ -field generated by events up to time τ_n .

Let $x(t) := (x_1(t), \dots, x_L(t))^T$ denote the vector of queue lengths. In accordance with terminology of Markov Decision Processes, we will call $x(t)$ the "state." A policy whose action at any time t depends only on $x(t)$ is called stationary, again in accordance with the terminology of Markov Decision Processes. Under a stationary policy, the system is described by the Markov chain $\{x(t)\}$.

We will treat a larger class of scheduling policies than stationary policies. We will consider any scheduling policy which takes a constant action in intervals of the form $[\tau_n, \tau_{n+1})$, and call such a policy noninterruptive. (The term noninterruptive should not be confused with the term nonpreemptive.) As

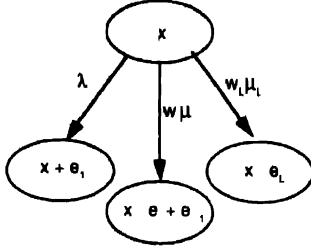


Fig. 2 State transition diagram for basic open re-entrant line

an example, the well-known FCFS policy is noninterruptive. Note that any scheduling policy that does not change actions between real transition epochs is noninterruptive. Of course, all stationary policies are of this form and are hence automatically noninterruptive.

We will allow preemptive priority at an epoch, if the scheduling policy calls for it.

Let $e_i = (0, \dots, 0, 1, 0, \dots, 0)^T$ be the i th unit vector. The state transition diagram of the network is as shown in Fig. 2 for any noninterruptive policy.

Let us consider the quadratic form $r^T(\tau_n)Qr(\tau_n)$. Note that since $r(\tau_n)$ can grow no faster than linearly in n , the conditional expectation $E[r^T(\tau_{n+1})Qr(\tau_{n+1}) | \mathcal{F}_\tau]$ exists. From Fig. 2, for any noninterruptive policy we obtain

$$\begin{aligned} E[r^T(\tau_{n+1})Qr(\tau_{n+1}) | \mathcal{F}_\tau] &= \lambda(r(\tau_n) + e_1)^T Q(r(\tau_n) + e_1) \\ &+ \sum_{i=1}^{L-1} \mu_i w_i(\tau_n)(r(\tau_n) - e_i + e_{i+1})^T Q(r(\tau_n) - e_i + e_{i+1}) \\ &+ \mu_L w_L(\tau_n)(r(\tau_n) - e_L)^T Q(r(\tau_n) - e_L) \\ &+ \sum_{i=1}^L \mu_i (1 - u_i(\tau_n)) r^T(\tau_n) Q r(\tau_n) \end{aligned} \quad (3)$$

Using (2) and the symmetry of Q as in (1) we obtain

$$\begin{aligned} E[r^T(\tau_{n+1})Qr(\tau_{n+1}) | \mathcal{F}_\tau] &= r^T(\tau_n)Qr(\tau_n) + 2\lambda e_1^T Q r(\tau_n) + \lambda e_1^T Q e_1 \\ &+ 2 \sum_{i=1}^{L-1} \mu_i w_i(\tau_n)(e_{i+1} - e_i)^T Q r(\tau_n) \\ &+ \sum_{i=1}^{L-1} \mu_i w_i(\tau_n)(e_{i+1} - e_i)^T Q (e_{i+1} - e_i) \\ &- 2\mu_L w_L(\tau_n) e_L^T Q r(\tau_n) + \mu_L w_L(\tau_n) e_L^T Q e_L \end{aligned}$$

Note now that since $w_i(\tau_n) = 0$ or 1 , all the terms not featuring $r(\tau_n)$ above are bounded above by a constant, i.e.,

$$\begin{aligned} \lambda e_1^T Q e_1 + \sum_{i=1}^{L-1} \mu_i w_i(\tau_n)(e_{i+1} - e_i)^T Q (e_{i+1} - e_i) \\ + \mu_L w_L(\tau_n) e_L^T Q e_L \leq M < +\infty \end{aligned}$$

Hence

$$\begin{aligned} E[r^T(\tau_{n+1})Qr(\tau_{n+1}) | \mathcal{F}_\tau] &\leq r^T(\tau_n)Qr(\tau_n) + 2\lambda e_1^T Q r(\tau_n) \\ &+ 2 \sum_{i=1}^{L-1} \mu_i w_i(\tau_n)(e_{i+1} - e_i)^T Q r(\tau_n) \\ &- 2\mu_L w_L(\tau_n) e_L^T Q r(\tau_n) + M \end{aligned}$$

Let us suppose that the initial condition is deterministic (or more generally, a bounded random variable or even more generally, has a finite second moment). As noted earlier, $r(\tau_n)$ grows no faster than linearly in n . Hence $E[r^T(\tau_n)Qr(\tau_n)]$ exists for every n . By taking the unconditional expectation we obtain

$$\begin{aligned} E[r^T(\tau_{n+1})Qr(\tau_{n+1})] &\leq E[r^T(\tau_n)Qr(\tau_n)] + 2\lambda e_1^T Q E[r(\tau_n)] \\ &+ 2 \sum_{i=1}^{L-1} \mu_i E[u_i(\tau_n)(e_{i+1} - e_i)^T Q r(\tau_n)] \\ &- 2\mu_L E[w_L(\tau_n) e_L^T Q r(\tau_n)] + M \end{aligned} \quad (4)$$

Let us denote

$$z_{ij}(\tau_n) = w_i(\tau_n) e_j^T Q r(\tau_n) \quad (5)$$

Using (5), (4) can be written as

$$\begin{aligned} E[r^T(\tau_{n+1})Qr(\tau_{n+1})] &\leq E[r^T(\tau_n)Qr(\tau_n)] \\ &+ 2\lambda \sum_{j=1}^L q_{1j} E[r_j(\tau_n)] \\ &+ 2 \sum_{i=1}^{L-1} \mu_i \sum_{j=1}^L (q_{i+1,j} - q_{ij}) E[r_j(\tau_n)] \\ &- 2\mu_L \sum_{j=1}^L q_{Lj} E[r_j(\tau_n)] + M \end{aligned}$$

By summing over n , and telescoping, we obtain

$$\begin{aligned} \frac{1}{N+1} \sum_{n=0}^N \left[\lambda \sum_{j=1}^L q_{1j} E[r_j(\tau_n)] \right. \\ \left. - \sum_{i=1}^{L-1} \mu_i \sum_{j=1}^L (q_{i+1,j} - q_{ij}) E[r_j(\tau_n)] \right. \\ \left. + \mu_L \sum_{j=1}^L q_{Lj} E[r_j(\tau_n)] \right] \\ \leq \frac{1}{2(N+1)} (E[r^T(0)Qr(0)] \\ - E[r^T(\tau_{N+1})Qr(\tau_{N+1})]) \\ + \frac{M}{2} \leq M' < +\infty \text{ for all } N \end{aligned} \quad (6)$$

In the last inequality above, we have used the nonnegativity of $r^T(\tau_{N+1})Qr(\tau_{N+1})$, which is guaranteed by the copositivity condition (1), since $r(\tau_{N+1})$ lies in the nonnegative orthant.

Now note that if we can find a $\gamma > 0$ so that

$$\begin{aligned} & \lambda \sum_{j=1}^L q_{1j} E(x_j(\tau_n)) \\ & + \sum_{i=1}^{L-1} \mu_i \sum_{j=1}^L (q_{i+1,j} - q_{ij}) E(z_{ij}(\tau_n)) \\ & - \mu_L \sum_{j=1}^L q_{Lj} E(z_{Lj}(\tau_n)) \leq -\gamma \sum_{j=1}^L E(x_j(\tau_n)) \end{aligned}$$

then from (6) we would have stability-in-the-mean, i.e.,

$$\frac{1}{N+1} \sum_{n=0}^N \sum_{j=1}^L E(x_j(\tau_n)) \leq M'' < +\infty \text{ for all } N. \quad (7)$$

Before we pursue the issue of finding such a γ , we point out certain consequences of stability-in-the-mean for stationary, nonidling policies. In the rest of this paper we will restrict attention to scheduling policies that are nonidling, i.e., whenever one of the buffers at a machine is nonempty, then the machine cannot stay idle. For stationary, nonidling policies, $\{x(\tau_n)\}$ is a time-homogeneous, countable state, Markov chain, which has a single communicating class that is aperiodic (since the origin can be reached from every state, and the system can stay at the origin for two consecutive time steps). The condition (7) then guarantees positive recurrence, i.e., the existence of a unique steady state probability distribution. To see this, note that if the chain is not positively recurrent, then the probability that the chain is in a fixed finite set of states converges to zero as $n \rightarrow \infty$. Then, that is so even for the finite set of states $\{x: \sum_{j=1}^L x_j \leq M'', \text{ and all } x_j \geq 0 \text{ and integral}\}$. This contradicts (7). Moreover, the Markov chain has bounded first moment, and the mean total number of customers converges to a finite steady state value. In fact, we will show in the next section that it even establishes the geometric convergence of an exponential moment.

Let us now see how to assure (7) for some $\gamma > 0$. We will actually work at assuring that the inequality (7) holds without the expectation being taken, i.e.,

$$\begin{aligned} & \lambda \sum_{j=1}^L q_{1j} x_j(\tau_n) + \sum_{i=1}^{L-1} \mu_i \sum_{j=1}^L (q_{i+1,j} - q_{ij}) z_{ij}(\tau_n) \\ & - \mu_L \sum_{j=1}^L q_{Lj} z_{Lj}(\tau_n) \leq -\gamma \sum_{j=1}^L x_j(\tau_n). \end{aligned} \quad (8)$$

Let us now motivate the reason for restricting our attention to nonidling policies in our tests for stability. Note that the coefficients λq_{1j} of the $x_j(\tau_n)$'s on the left-hand side (LHS) above are all nonnegative, while the corresponding coefficient $(-\gamma)$ on the right-hand side (RHS) is negative. Clearly, to assure (8) it is necessary that there exist some choice of constants $\{\alpha_{ij}\}$ for which

$$\sum_{j=1}^L x_j(\tau_n) \leq \sum_{i=1}^L \sum_{j=1}^L \alpha_{ij} z_{ij}(\tau_n).$$

Focusing on a fixed index j , one will in particular need $x_j(\tau_n)$ to be bounded above by some linear combination of

$\{z_{ij}(\tau_n): 1 \leq i \leq L\}$. This can only be assured if some machine is guaranteed to be working whenever $x_j > 0$; hence the restriction to nonidling scheduling policies in the sequel.

Let us return to (8). For notational convenience, let us define

$$q_{L+1,j} := 0 \text{ for all } j = 1, 2, \dots, L.$$

Focusing on a fixed value of the index j , it is clear that (8) is assured, if

$$\begin{aligned} & \lambda q_{1j} x_j(\tau_n) + \sum_{i=1}^L \mu_i (q_{i+1,j} - q_{ij}) z_{ij}(\tau_n) \\ & \leq -\gamma x_j(\tau_n) \text{ for all } j = 1, 2, \dots, L. \end{aligned} \quad (9)$$

Grouping the terms by machine, i.e., using $\{1, 2, \dots, L\} = \bigcup_{\sigma} \{i: \sigma(i) = \sigma\}$, we see that the LHS of (9) satisfies

$$\begin{aligned} & \lambda q_{1j} x_j(\tau_n) + \sum_{i=1}^L \mu_i (q_{i+1,j} - q_{ij}) z_{ij}(\tau_n) \\ & = \lambda q_{1j} x_j(\tau_n) + \sum_{\sigma=1}^S \sum_{\{i: \sigma(i)=\sigma\}} \mu_i (q_{i+1,j} - q_{ij}) z_{ij}(\tau_n) \\ & \leq \lambda q_{1j} x_j(\tau_n) + \sum_{\sigma=1}^S \max_{\{i: \sigma(i)=\sigma\}} \mu_i (q_{i+1,j} - q_{ij}) \\ & \quad \times \sum_{\{i: \sigma(i)=\sigma\}} z_{ij}(\tau_n) \\ & = \lambda q_{1j} x_j(\tau_n) + \max_{\{i: \sigma(i)=\sigma(j)\}} \mu_i (q_{i+1,j} - q_{ij}) \\ & \quad \times \sum_{\{i: \sigma(i)=\sigma(j)\}} z_{ij}(\tau_n) + \sum_{\{\sigma': \sigma' \neq \sigma(j)\}} \max_{\{i: \sigma(i)=\sigma'\}} \mu_i (q_{i+1,j} - q_{ij}) \\ & \quad \times \sum_{\{i: \sigma(i)=\sigma'\}} z_{ij}(\tau_n). \end{aligned} \quad (10)$$

We now investigate how to assure that the RHS of (10) can be bounded above by $-\gamma x_j(\tau_n)$. For nonidling policies

$$\sum_{\{i: \sigma(i)=\sigma\}} x_i(\tau_n) > 0 \Rightarrow \sum_{\{i: \sigma(i)=\sigma\}} w_i(\tau_n) = 1$$

and so

$$x_j(\tau_n) = \sum_{\{i: \sigma(i)=\sigma(j)\}} w_i(\tau_n) x_j(\tau_n).$$

Hence

$$x_j(\tau_n) = \sum_{\{i: \sigma(i)=\sigma(j)\}} z_{ij}(\tau_n). \quad (11)$$

Moreover, since other machines need not be working when b_j is nonempty, we have the nonidling inequalities

$$x_j(\tau_n) \geq \sum_{\{i: \sigma(i)=\sigma'\}} z_{ij}(\tau_n) \text{ for all } \sigma' \neq \sigma(j). \quad (12)$$

Employing (11), (12), in (10), we obtain

$$\begin{aligned} & \lambda q_{1j} + \sum_{i=1}^L \mu_i (q_{i+1,j} - q_{ij}) z_{ij}(\tau_n) \\ & \leq \lambda q_{1j} + \sum_{\{\sigma : \sigma(i) = \sigma(j)\}} \left[\max_{\{\sigma(i) = \sigma(j)\}} \mu_i (q_{i+1,j} - q_{ij}) \right] z_{ij}(\tau_n) \\ & \quad + \sum_{\{\sigma : \sigma \neq \sigma(j)\}} \left[\max_{\{\sigma(i) = \sigma\}} \mu_i (q_{i+1,j} - q_{ij}) \right]^+ z_{ij}(\tau_n) \end{aligned}$$

Above, $[y]^+ = \max\{y, 0\}$ denotes the positive part of y . The sign needs to be taken into account, since a negative sign of $\mu_i (q_{i+1,j} - q_{ij})$ may reverse the inequality.

Hence we see that if we can find an appropriate set of $\{q_{ij}\}$ for which

$$\begin{aligned} & \lambda q_{1j} + \max_{\{\sigma(i) = \sigma(j)\}} \mu_i (q_{i+1,j} - q_{ij}) \\ & \quad + \sum_{\{\sigma : \sigma \neq \sigma(j)\}} \left[\max_{\{\sigma(i) = \sigma\}} \mu_i (q_{i+1,j} - q_{ij}) \right]^+ \\ & \leq -\gamma \quad \text{for } j = 1, 2, \dots, I \end{aligned}$$

for some $\gamma > 0$, then (8), and hence (7), is assured. Thus we have arrived at the following theorem.

Theorem 1—The Basic Characterization. Consider the basic open re-entrant line. Suppose there exists asymmetric copositive matrix $Q = [q_{ij}]$ which satisfies the following conditions. For every $j = 1, 2, \dots, I$

$$\begin{aligned} & \lambda q_{1j} + \max_{\{\sigma(i) = \sigma(j)\}} \mu_i (q_{i+1,j} - q_{ij}) \\ & \quad + \sum_{\{\sigma : \sigma \neq \sigma(j)\}} \left[\max_{\{\sigma(i) = \sigma\}} \mu_i (q_{i+1,j} - q_{ij}) \right]^+ < 0 \quad (13) \end{aligned}$$

(Above $q_{I+1,j} = 0$ for all j .) Then, every nonidling, noninterruptive scheduling policy is stable in the mean, i.e., there exist constants ϵ, C' such that

$$\frac{1}{N} \sum_{n=0}^N \sum_{j=1}^I E(z_{ij}(\tau_n)) \leq \frac{\epsilon F(z^I(0)Qz(0))}{N} + C' \quad \text{for all } N$$

Moreover, if the scheduling policy is stationary, then there is a unique steady state probability distribution.

V FROM STABILITY TO GEOMETRIC CONVERGENCE OF AN EXPONENTIAL MOMENT

In fact, for a Markovian system, the above Lyapunov based negative drift argument actually establishes the geometric convergence of an exponential moment (defined below). Thus, in particular it establishes the finiteness of all (polynomial) moments and their geometric ergodicity.

To see this, we simply work with the square root of the earlier Lyapunov function. From Fig. 2, just as we obtained (3), we obtain

$$\begin{aligned} & L \left[\sqrt{z^I(\tau_{n+1})Qz(\tau_{n+1})} \middle| \mathcal{F}_{(\tau_n)} \right] \\ & = \lambda \sqrt{(z(\tau_n) + e_1)^T Q(z(\tau_n) + e_1)} \end{aligned}$$

$$\begin{aligned} & + \sum_{i=1}^{I-1} \mu_i w_i(\tau_n) \\ & \times \sqrt{(z(\tau_n) - e_i + e_{i+1})^T Q(z(\tau_n) - e_i + e_{i+1})} \\ & + \mu_L w_L(\tau_n) \sqrt{(z(\tau_n) - e_I)^T Q(z(\tau_n) - e_I)} \\ & + \sum_{i=1}^L \mu_i (1 - u_i(\tau_n)) \sqrt{z^I(\tau_n)Qz(\tau_n)} \quad (14) \end{aligned}$$

From the concavity of the square root we obtain

$$\begin{aligned} & L \left[\sqrt{z^I(\tau_{n+1})Qz(\tau_{n+1})} \middle| \mathcal{F}_{(\tau_n)} \right] \\ & \leq [z^I(\tau_n)Qz(\tau_n) + 2\lambda z^I(\tau_n)Qz(\tau_n) \\ & \quad + 2 \sum_{i=1}^{I-1} \mu_i w_i(\tau_n) (e_{i+1} - e_i)^T Qz(\tau_n) \\ & \quad - 2\mu_L w_L(\tau_n) e_I^T Qz(\tau_n) + M]^1/2 \quad (15) \end{aligned}$$

Now suppose the conditions of Theorem 1 are met. Then, from (8), the RHS above can be bounded as

$$\text{RHS of (15)} \leq \left[z^I(\tau_n)Qz(\tau_n) - 2\gamma \sum_{j=1}^I z_{ij}(\tau_n) + M \right]^1/2 \quad (16)$$

Now from Theorem 14.2.2 of [28] by taking τ there to be the first hitting time of the origin it follows that there exists a $\delta > 0$, such that $z^I Q z > \delta \|z\|^2$ for all large enough $\|z\|$ in the positive orthant.¹ (Or in the case where all $q_{ij} \geq 0$, the fact that $q_{ii} > 0$ follows trivially from inequality (iii) of Theorem 1.) Hence, there exists an $\epsilon > 0$ small enough so that

$$\text{RHS of (16)} \leq [z^I(\tau_n)Qz(\tau_n)]^1/2 - \epsilon$$

whenever

$$\sum_{j=1}^I z_{ij}(\tau_n) > M'''$$

for some large M''' .

Letting $W(\tau_n) = \sqrt{z(\tau_n)^T Q z(\tau_n)}$ we have thus shown that

$$E[W(\tau_{n+1}) | \mathcal{F}_{(\tau_n)}] \leq W(\tau_n) - \epsilon \quad \text{if } z(\tau_n) \text{ lies outside a compact set,}$$

and is bounded when $z(\tau_n)$ is in the compact set. Moreover, the state can jump by only a bounded amount at each transition, and hence $W(\tau_{n+1}) - W(\tau_n)$ is bounded. From these two facts, it follows that the Markov chain representing the evolution of the system has a geometrically converging exponential moment, see [28, Theorem 16.3.1].

¹Thus Q is actually strictly copositive.

Theorem 2—Geometric Convergence of an Exponential Moment: Consider the basic, open re-entrant line. Suppose that the scheduling policy is stationary and that all the conditions of Theorem 1 are satisfied. Then the Markov chain $\{x(\tau_n)\}$ has a geometrically converging exponential moment,² i.e., there exist $\epsilon > 0$, $r > 1$, and $C < \infty$, such that for any function f satisfying $|f(y)| \leq \exp(\epsilon\|y\|)$ for all y , and any initial condition $x(\tau_0) = x$,

$$\sum_{n=0}^{\infty} r^n |E[f(x(\tau_n))]| - \sum_y f(y) \pi(y) < C \exp(\epsilon\|x\|) \text{ for all } x.$$

Above, $\pi(y)$ denotes the steady-state probability of the state y . Hence, in particular, the Markov chain admits a finite exponential moment. That is, for some $C' < \infty$

$$E[\exp(\epsilon\|x(\tau_n)\|)] \leq C' \exp(\epsilon\|x\|) < \infty \text{ for all } n.$$

The reader is referred to Meyn and Tweedie [29] for estimates of the rate of convergence. It is worth mentioning that the uniformization procedure is just a way of computing the drift $\mathcal{A}^T Qx$, where \mathcal{A} is the extended generator for the unsampled Markov process. Thus one actually has, for some $\rho < 1$, $|E[f(x_t)] - \sum_y f(y) \pi(y)| < C\rho^t V(x)$ for all x and all $t \geq 0$, i.e., a similar geometric convergence for the original unsampled chain.

VI. A LINEAR PROGRAMMING CHARACTERIZATION

As noted earlier, if Q is a symmetric nonnegative matrix, then it is copositive. Note now that the LHS of (13) in Theorem 1 is homogeneous in Q . Hence, if (13) is valid, then by multiplying Q by arbitrarily large positive numbers, one can drive the value of the LHS of the inequality (13) in Theorem 1 to $-\infty$. This allows us to provide a sufficient condition for stability in terms of the unboundedness of a linear program.

Theorem 3—A Linear Programming Characterization: Consider the basic, open re-entrant line. Suppose that the following linear program has an unbounded solution

$$\text{Max } \gamma$$

subject to the constraints

$$\lambda q_{1j} + r_j + \sum_{\{\sigma' : \sigma' \neq \sigma(j)\}} s_{\sigma',j} + \gamma \leq 0 \quad \text{for all } j$$

$$r_j \geq \mu_i(q_{i+1,j} - q_{i,j}) \quad \text{for all } i \in I(j), \text{ and for all } j$$

$$s_{\sigma,j} \geq \mu_i(q_{i+1,j} - q_{i,j}) \quad \text{for all } i \text{ with } \sigma(i) = \sigma, \text{ and all } j$$

$$q_{i,j} = q_{j,i} \quad \text{for all } i, j$$

$$q_{L+1,j} = 0 \quad \text{for all } j$$

$$q_{i,j} \geq 0 \quad \text{for all } i, j$$

$$s_{\sigma,j} \geq 0 \quad \text{for all } \sigma, j$$

r_j unrestricted in sign.

Then, every nonidling, noninterruptive policy is stable in the mean. Moreover, every nonidling, stationary policy has a geometrically converging exponential moment.

²This property is called " $\exp(\epsilon\|x\|)$ -uniform ergodicity" in [28].

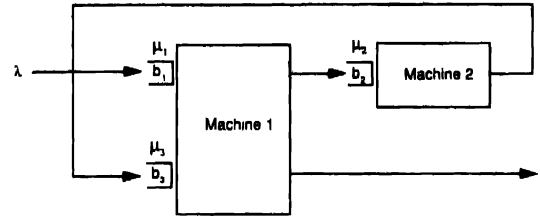


Fig. 3. Example of Sections VII and X

The number of variables $\{q_{ij}, r_k, s_{\sigma j}, \gamma\}$ in the above linear program is $\frac{L(L+1)}{2} + L + (S-1)L + 1$. (Note that the variables $s_{\sigma(j),j}$, $q_{i,j}$ with $i > j$, and $q_{L+1,j}$ are not really needed.) The number of constraints is $L + \sum_{j=1}^L |I(j)| + \sum_{j=1}^L \sum_{\sigma \neq \sigma(j)} |\{i : \sigma(i) = \sigma\}|$. It may be possible to rewrite the linear program more economically.

We find it convenient to slightly modify the linear program in Theorem 3 by bounding γ by 1. Thus if the value of the linear program is 1, then one deduces stability (rather than from the unboundedness of the value as in Theorem 3).

Corollary 1—A 0-1 Test of Stability: Consider the same linear program as in Theorem 3, except that we impose the additional constraint

$$\gamma \leq 1.$$

If the linear program has value one, then every nonidling noninterruptive policy is stable-in-the-mean. Moreover, every nonidling, stationary policy has a geometrically converging exponential moment. If the value of the linear program is zero, however, then no conclusion can be drawn regarding stability or instability.

VII. EXAMPLE: ALL NONIDLING POLICIES STABLE

Consider the system shown in Fig. 3. Then, to show that there exists a $Q = Q^T$ satisfying (7), it is sufficient to find $q_{i,j} = q_{j,i} \geq 0$, so that

$$\begin{aligned} & [\lambda q_{11}, \lambda q_{12}, \lambda q_{13}] \begin{bmatrix} x_1(\tau_n) \\ x_2(\tau_n) \\ x_3(\tau_n) \end{bmatrix} \\ & + [-\mu_1 q_{11} + \mu_1 q_{12}, -\mu_1 q_{12} + \mu_1 q_{22}, -\mu_1 q_{13} + \mu_1 q_{23}, \\ & \quad -\mu_2 q_{12} + \mu_2 q_{13}, -\mu_2 q_{22} + \mu_2 q_{23}, -\mu_2 q_{23} + \mu_2 q_{33}, \\ & \quad -\mu_3 q_{13}, -\mu_3 q_{23}, -\mu_3 q_{33}] \begin{bmatrix} z_{11}(\tau_n) \\ z_{12}(\tau_n) \\ z_{13}(\tau_n) \\ z_{21}(\tau_n) \\ z_{22}(\tau_n) \\ z_{23}(\tau_n) \\ z_{31}(\tau_n) \\ z_{32}(\tau_n) \\ z_{33}(\tau_n) \end{bmatrix} \\ & \leq -\gamma[x_1(\tau_n) + x_2(\tau_n) + x_3(\tau_n)]. \end{aligned}$$

By the nonidling condition, $x_1(\tau_n) = z_{11}(\tau_n) + z_{31}(\tau_n)$, $x_2(\tau_n) = z_{22}(\tau_n)$ and $x_3(\tau_n) = z_{13}(\tau_n) + z_{33}(\tau_n)$. We thus see that it suffices to show that one can choose $\{q_{11}, q_{12}, q_{13}, q_{22}, q_{23}, q_{33}\}$, all nonnegative, so that $\lambda q_{11} - \mu_1 q_{11} + \mu_1 q_{12} < 0$, $-\mu_1 q_{12} + \mu_1 q_{22} < 0$, $-\mu_1 q_{13} + \mu_1 q_{23} +$

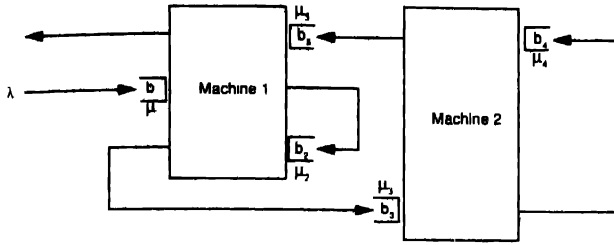


Fig. 4 Dai Wang example of Section VIII

$$\begin{aligned} \lambda q_{11} < 0, -\mu_2 q_{12} + \mu_2 q_{13} < 0, -\mu_2 q_{22} + \mu_2 q_{23} + \lambda q_{12} < 0 \\ \mu_2 q_{23} + \mu_2 q_{33} < 0, -\mu_3 q_{13} + \lambda q_{11} < 0, -\mu_3 q_{23} < 0, \\ \mu_3 q_{33} + \lambda q_{11} < 0 \end{aligned}$$

Let us suppose that $\mu_1 = \mu_2 = \mu_3 = \mu$, and $\rho = \frac{2\lambda}{\mu}$. Then the above is equivalent to, $(1 - \rho/2)q_{11} > q_{12}$, $q_{12} > q_{22}$, $(1 - \rho/2)q_{13} > q_{23}$, $q_{12} > q_{13}$, $q_{22} > \rho q_{12}/2 + q_{23}$, $q_{24} > q_{13}$, $q_{13} > \rho q_{11}/2$, $q_{24} > 0$, $q_{33} > \rho q_{13}/2$.

It is easily checked that if $\rho < 1$ then one can choose $q_{ij} > 0$ which satisfy the above conditions.

Hence, we conclude that if $\mu_1 = \mu_2 = \mu_3 = \mu$ and $\rho = \frac{2\lambda}{\mu} < 1$, then all nonidling, noninterruptive, scheduling policies are stable in the mean. This includes the FCFS policy, which is already known to be stable since $\mu_1 = \mu_3$ see Kelly [30]. Moreover, every nonidling stationary policy has a geometrically converging exponential moment.

VIII THE DAI WANG EXAMPLE

Dai and Wang [12] (see also [14]) show that the system of Fig. 4 does not have a Brownian approximation. It is a basic open re-entrant line with service rates $\mu_1 = 10$, $\mu_2 = 20$, $\mu_3 = 10/9$, $\mu_1 = 20$, and $\mu_3 = 5/1$. Let $\rho = \lambda/\mu_1 + \lambda/\mu_2 + \lambda/\mu_3 = \lambda/10 + \lambda/20 + 9\lambda/10$ be the load factor on the two machines. Our goal is to determine whether the system is stable for all $\rho < 1$.

First note from the equations involving λ in (13) and Corollary 1 that if the value of the linear program is one for some λ' then it is one for all $\lambda < \lambda'$. Hence there will be critical value λ_{crit} , such that the linear program has value one for $\lambda < \lambda_{crit}$ and value zero for $\lambda > \lambda_{crit}$. Equivalently, there exists such a ρ_{crit} . So we wish to see if $\rho_{crit} = 1$.

Investigating the linear program from Corollary 1, we find that its value is one for $\lambda < 0.55587$ (approximately) and zero for $0.55587 < \lambda < 1$. Thus we can only assert stability for $\rho < 0.95(0.55587) = 0.528$.

IX BUFFER PRIORITY POLICIES

Consider the basic, open re-entrant line. Suppose that at every machine σ there is a rank ordering of the set of buffers $\{b_i : \sigma(i) = \sigma\}$ served by the machine according to which preemptive priority is given by the machine. To describe such a buffer priority policy more concisely, let $\{\theta(1) \dots \theta(L)\}$ be a permutation of $\{1, 2, \dots, L\}$, with preference given to b_i over b_j if $\theta(i) < \theta(j)$ and both b_i and b_j share the same machine, i.e. $\sigma(i) = \sigma(j)$. The policy is nonidling, stationary, and preemptive.

Then, $x_j(\tau_n) > 0 \Rightarrow w_j(\tau_n) = 0$, if $\theta(i) < \theta(j)$ and $\sigma(i) = \sigma(j)$. Hence $z_{ji} = w_j(\tau_n)z_i(\tau_n) = 0$, if $\theta(i) < \theta(j)$

and $\sigma(i) = \sigma(j)$. As a consequence

$$\begin{aligned} x_j(\tau_n) &= \sum_{\{i : i \in I(j), \theta(i) \leq \theta(j)\}} z_{ij}(\tau_n) \\ &\geq \sum_{\{i : \sigma(i) = \sigma\}} z_{ij}(\tau_n) \text{ for } \sigma' \neq \sigma \end{aligned}$$

Recall from Section IV that our goal is to determine symmetric copositive Q satisfying

$$\begin{aligned} \lambda q_{1j} r_j(\tau_n) + \sum_{i \in I(j)} \mu_i (q_{i+1,j} - q_{ij}) z_{ij}(\tau_n) \\ + \sum_{\{\sigma : \sigma \neq \sigma(j)\}} \sum_{\{i : \sigma(i) = \sigma\}} \mu_i (q_{i+1,j} - q_{ij}) z_{ij}(\tau_n) \\ \leq -\gamma r_j(\tau_n) \text{ for } j = 1, 2, \dots, L \end{aligned}$$

for some $\gamma > 0$. (This can be seen by rewriting the LHS of (9) as in the first equality of (10)). Using (17), (18) the conditions required to establish stability in Corollary 1 can be relaxed.

Theorem 4—Stability of Buffer Priority Policies. Consider the basic open re-entrant line. Let $\{\theta(1) \dots \theta(L)\}$ be a permutation of $\{1, 2, \dots, L\}$. Consider the preemptive buffer priority policy which gives preference to b_i over b_j if $\theta(i) < \theta(j)$ and both share the same machine. Then the buffer priority policy $\theta(\cdot)$ is stable, with a geometrically converging exponential moment, if the following linear program has value one.

$$\text{Max } \gamma$$

subject to

$$\begin{aligned} \lambda q_{1j} + \max_{\{i : i \in I(j) \text{ and } \theta(i) \leq \theta(j)\}} \mu_i (q_{i+1,j} - q_{ij}) \\ + \sum_{\{\sigma : \sigma \neq \sigma(j)\}} \left[\max_{\{i : \sigma(i) = \sigma\}} \mu_i (q_{i+1,j} - q_{ij}) \right] + \gamma \\ \leq 0 \text{ for } j = 1, 2, \dots, L \\ \gamma \leq 1 \\ q_{ij} - q_{ji} \geq 0 \text{ for } 1 \leq i, j \leq L \end{aligned}$$

X EXAMPLE LBFS AND FBFS ARE STABLE

Consider the system of Fig. 3. Unlike in Section VII we do not require that all the μ_i 's are equal.

First let us examine the class of nonidling policies. Can we prove that all nonidling, noninterruptive policies are stable? To investigate, we consider the following linear program

$$\text{Max } \gamma \quad (19)$$

subject to the constraints

$$\begin{aligned} \lambda q_{11} + \max\{\mu_1(q_{12} - q_{11}) - \mu_3 q_{13}\} \\ + \max\{\mu_2(q_{13} - q_{12}), 0\} + \gamma \leq 0 \end{aligned} \quad (20)$$

$$\begin{aligned} \lambda q_{12} + \mu_2(q_{23} - q_{22}) + \max\{\mu_1(q_{22} - q_{12}) \\ - \mu_3 q_{23}, 0\} + \gamma \leq 0 \end{aligned} \quad (21)$$

$$\begin{aligned} \lambda q_{13} + \max\{\mu_1(q_{23} - q_{13}) - \mu_3 q_{33}\} \\ + \max\{\mu_2(q_{33} - q_{23}), 0\} + \gamma \leq 0 \end{aligned} \quad (22)$$

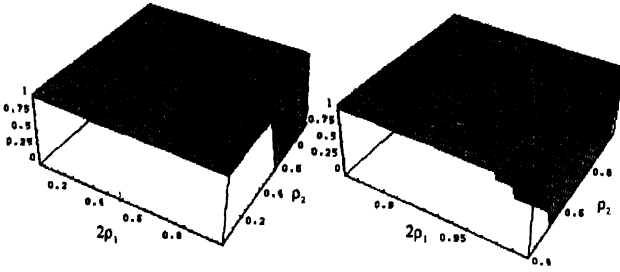


Fig. 5. Value of LP for all nonidling, noninterruptive policies in example of Section X. The figure on the right is a more detailed view of one corner of the figure on the left.

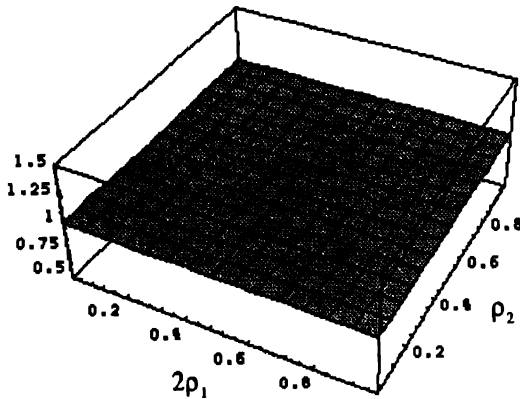


Fig. 6. Value of LP for example of Section X for the LBFS policy and for the FBFS policy.

$$\gamma \leq 1 \quad (23)$$

$$q_{11}, q_{12}, q_{13}, q_{22}, q_{23}, q_{33} \geq 0. \quad (24)$$

Let $\rho_i := \frac{\lambda_i}{\mu_i}$, and consider $\rho_1 = \rho_3$. Fig. 5 plots the value of the linear program as a function of $2\rho_1$ and ρ_2 . It shows that there is a region where the value of the linear program is 0, and for such values of $\rho_1, \rho_2, \rho_3 (= \rho_1)$, no conclusion regarding stability or instability can be drawn.

Now consider the LBFS policy. The corresponding linear program is the same as (19)–(24), with exception that (22) is changed to

$$\lambda q_{13} - \mu_3 q_{33} + \max\{\mu_2(q_{33} - q_{23}), 0\} + \gamma \leq 0. \quad (25)$$

The plot of its value as a function of $2\rho_1$, and ρ_2 is given in Fig. 6. It shows that the system is stable for all values of $2\rho_1 < 1$ and $\rho_2 < 1$. Hence, for $\rho_1 = \rho_3$, the LBFS policy is stable in the entire capacity region.

Now turn to the FBFS policy. Its linear program is also almost the same as (19)–(24), except that (20) is changed to the following

$$\lambda q_{11} + \mu_1(q_{12} - q_{11}) + \max\{\mu_2(q_{13} - q_{12}), 0\} + \gamma \leq 0. \quad (26)$$

The plot of its value is the same as that of LBFS, as shown in Fig. 6. Hence it is also stable in the entire stability region, when $\rho_1 = \rho_3$.

XI. A NONLINEAR PROGRAMMING CHARACTERIZATION

Note that every symmetric positive semidefinite Q is copositive. Thus, in our stability tests, we could use the class of symmetric positive semidefinite matrices. Every such Q possesses a square root A , i.e., $Q = A^T A$. Hence, one may search over the unrestricted space of a_{ij} 's, rather than the space of q_{ij} 's. This yields the following Theorem.

Theorem 5—A Nonlinear Programming Characterization. Consider the basic, open, re-entrant line. Consider the nonlinear program

$$\text{Max } \gamma$$

subject to all the constraints of Corollary 1, except that every q_{ij} is replaced by $\sum_{k=1}^L a_{ki} a_{kj}$, and the nonnegativity constraint on the q_{ij} 's is removed. If this nonlinear program has value one, then, every nonidling, noninterruptive, scheduling policy is stable in the mean. Moreover, then, every nonidling, stationary policy gives rise to a Markov chain, with a geometrically converging exponential moment.

Also, one can extend this Theorem to search over convex combinations of a nonnegative matrix, and a positive semidefinite matrix. This yields the following theorem.

Theorem 6—A More General Nonlinear Programming Characterization Consider the basic, open, re-entrant line. Consider the nonlinear program

$$\text{Max } \gamma$$

subject to all the constraints of Corollary 1, except that every q_{ij} is replaced by $q'_{ij} + \sum_{k=1}^L a_{ki} a_{kj}$, and the nonnegativity constraint on the q_{ij} 's is replaced by nonnegativity constraints on the q'_{ij} 's, while the a_{ij} 's are unrestricted. If this nonlinear program has value one, then every nonidling, noninterruptive, scheduling policy is stable in the mean. Moreover, then, every nonidling, stationary policy gives rise to a Markov chain, with a geometrically converging exponential moment.

Both these theorems can be extended in the same ways as Theorem 1, to treat various kinds of systems and scheduling policies.

To go beyond Theorems 5 and 6 and obtain the most powerful test obtainable through our approach, one could simply check whether the value of the linear program in Theorem 3 is one, without imposing any sign restrictions on q_{ij} , i.e., after removing the constraints $q_{ij} \geq 0$. If the value is indeed one, one can then test whether the obtained Q is copositive, using an algorithm as in [25]. As noted in Section III, however, the test of copositivity is NP-Complete and may therefore be computationally complex for systems of large size.

XII. MORE GENERAL QUEUEING NETWORKS

Consider a queueing network with S machines and L buffers. Let us suppose every buffer b_i has an exogenous Poisson arrival process of rate λ_i . A part leaving buffer b_i goes to buffer b_j with probability p_{ij} and leaves the system with probability $(1 - \sum_{j=1}^L p_{ij})$. The service time of parts in buffer b_i is exponential with mean $\frac{1}{\mu_i}$. Let us rescale time so

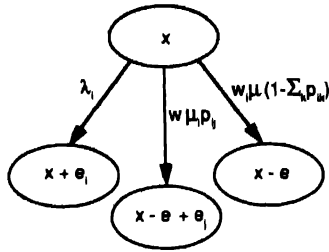


Fig 7 State transition diagram for general queueing network of Section XII

that $\sum_{i=1}^I \lambda_i + \sum_{j=1}^L \mu_j = 1$. The state transition diagram is shown in Fig 7. Hence

$$\begin{aligned} & P'(I^T(\tau_{n+1})Q(I(\tau_{n+1}) | \mathcal{F}(\tau_n))) \\ &= \sum_{i=1}^L \lambda_i (I(\tau_n) + e_i)^T Q(I(\tau_n) + e_i) \\ &+ \sum_{i=1}^I \mu_i w_i(\tau_n) \sum_{j=1}^L p_{ij} (I(\tau_n) - e_i + e_j)^T \\ &\times Q(I(\tau_n) - e_i + e_j) \\ &+ \sum_{i=1}^L \mu_i w_i(\tau_n) \left(1 - \sum_{j=1}^L p_{ij}\right) (I(\tau_n) - e_i)^T \\ &\times Q(I(\tau_n) - e_i) + \sum_{i=1}^I \mu_i (1 - w_i(\tau_n)) I^T(\tau_n) Q(I(\tau_n)) \end{aligned}$$

Proceeding as in Section IV, one may obtain the following theorem.

Theorem 7—Stability of a General Network. Consider the more general queueing network above operated under any nonidling, noninterruptive policy. Then it is stable in the mean, provided there exists a symmetric copositive Q that satisfies the following inequalities for $j = 1, 2, \dots, L$:

$$\begin{aligned} & \sum_{i=1}^I \lambda_i q_{ij} + \max_{\{\sigma: \sigma(i) = \sigma(j)\}} \mu_i \left(-q_{ij} + \sum_{k=1}^L p_{ik} q_{kj} \right) \\ &+ \sum_{\{\sigma: \sigma \neq \sigma(j)\}} \left[\max_{\{\sigma: \sigma(i) = \sigma\}} \mu_i \left(-q_{ij} + \sum_{k=1}^L p_{ik} q_{kj} \right) \right]^+ < 0 \end{aligned}$$

This can also be written as a linear program if we restrict $q_{ij} \geq 0$ for all i, j , or as a nonlinear program if we take $Q = A^T A + [q'_{ij}]$, with nonnegative q'_{ij} 's. Moreover, if the policy is stationary, then the system is a Markov chain with a geometrically converging exponential moment.

XIII THE RYBKO-STOLYAR EXAMPLE

Consider the system shown in Fig 8. The arrivals to buffers b_1 and b_3 are Poisson of rate λ . The service times are all exponentially distributed, with mean $1/\mu_1$ at buffers b_1 and b_3 , and mean $1/\mu_2$ at buffers b_2 and b_4 . Consider the buffer priority policy with ordering $\{b_4, b_2, b_3, b_1\}$, i.e., with priority given to buffers earlier in this list. Let us define $\rho_1 = \lambda/\mu_1$, and $\rho_2 = \lambda/\mu_2$.

Rybko and Stolyar [4] have shown that this system is unstable for $\lambda = 1$, if $\rho_2 > 1/2$ and $\rho_1 > 0$, even if

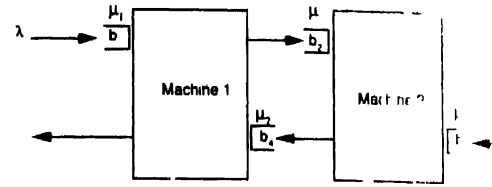


Fig 8 The Rybko-Stolyar example of Section XIII

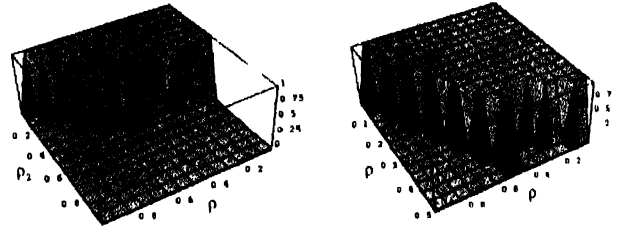


Fig 9 Value of the linear program for Rybko-Stolyar example of Section XIII. The figure on the right is a more detailed view of one portion of the figure on the left.

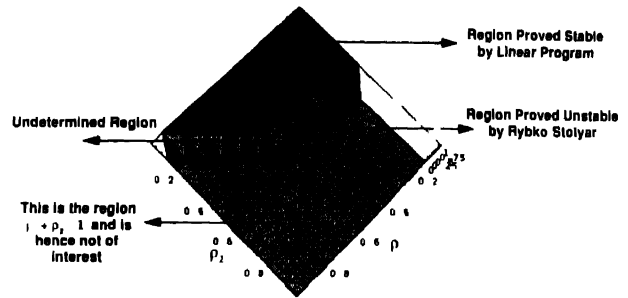


Fig 10 Stable, unstable and undetermined regions for Rybko-Stolyar example.

the service time requirements meet the capacity condition, $1/\mu_1 + 1/\mu_2 < 1$.

The following linear program tests the stability of the system for all ρ_1 and ρ_2 :

$$\text{Max } \gamma$$

subject to the constraints

$$\begin{aligned} & \text{Max}[\lambda q_{11} + \lambda q_{13} - \mu_1 q_{11} + \mu_1 q_{12} - \lambda q_{11} + \lambda q_{13} - \mu_1 q_{11}] \\ &+ \text{Max}[0 - \mu_3 q_{13} + \mu_3 q_{11}] < \gamma \end{aligned}$$

$$\lambda q_{12} + \lambda q_{23} - \mu_2 q_{22} + \text{Max}[0 - \mu_1 q_{11} + \mu_1 q_{22}] \leq \gamma$$

$$\begin{aligned} & \text{Max}[\lambda q_{13} + \lambda q_{33} - \mu_2 q_{23} - \lambda q_{13} + \lambda q_{33} - \mu_3 q_{33} + \mu_3 q_{13}] \\ &+ \text{Max}[0 - \mu_1 q_{13} + \mu_1 q_{23}] < \gamma \end{aligned}$$

$$\lambda q_{14} + \lambda q_{34} - \mu_1 q_{14} + \text{Max}[0 - \mu_3 q_{34} + \mu_3 q_{14}] \leq \gamma$$

$$\gamma < 1$$

$$q_{ij} \geq 0, \gamma \geq 0$$

Fig 9 plots the value of the linear program for $0 < \rho_1 < 1$ and $0 < \rho_2 < 1$. (The region $\rho_1 + \rho_2 \geq 1$ should be disregarded as it lies outside the capacity region.)

The following points are salient. First, the value of the linear program is zero in the region $\rho_2 > 1/2$ and thus noncontradictory with Rybko and Stolyar's result. Second, as shown in

Fig. 10, for most of the rest of the capacity region, the system is stable, since the value of the linear program is one. There is a small region, however, where the value of the linear program is zero; thus the stability remains unresolved there.

XIV. CONCLUDING REMARKS

We have provided here a programmatic procedure for establishing stability of queueing networks and scheduling policies.

There are several interesting questions which arise. First, it would be useful to study the structure of the linear or nonlinear programs and thus directly establish the stability of policies. We have done so analytically for the example of Section VII. Second, in all the examples tested by us, any Q giving a negative drift was always found to be a nonnegative matrix. It would be useful to determine whether there exists an example of a system where Q is copositive but not nonnegative. This should show that the more powerful tests of Section XI are in fact valuable. Third, it would be useful to implement a multi-step drift version of the above results. Finally, it would be useful to carry out a similar development for "instability" results, as in Fayolle [16] and Coffman *et al.*, [20].

ACKNOWLEDGMENT

The authors thank D. Down, S. Gowda, and J. Tsitsiklis for their useful comments.

REFERENCES

- [1] S. H. Lu and P. R. Kumar, "Distributed scheduling based on due dates and buffer priorities," *IEEE Trans Automat Contr*, vol. 36, pp. 1406–1416, Dec. 1991.
- [2] P. R. Kumar and T. I. Seidman, "Dynamic instabilities and stabilization methods in distributed real time scheduling of manufacturing systems," *IEEE Trans Automat Contr*, vol. 35, pp. 289–298, Mar. 1990.
- [3] C. J. Chase and P. J. Ramadge, "On the real time control of flexible manufacturing systems," in *Proc IEEE 28th Conf Decis Contr*, Tampa, FL, 1989, pp. 2026–2027.
- [4] A. N. Rybko and A. L. Stolyar, "On the ergodicity of stochastic processes describing open queueing networks," *Problemy Peredachi Informatsii*, vol. 28, pp. 2–26, 1991.
- [5] T. I. Seidman, "First come first serve is unstable," Univ. Maryland Baltimore County Tech. Rep. 1993.
- [6] M. Bramson, "Instability of FIFO queueing networks," Mathematics Dept. Univ. Wisconsin, Madison Tech. Rep. 1993.
- [7] J. M. Harrison, "Brownian models of heterogeneous customer populations," in *Stochastic Differential Systems: Stochastic Control Theory and Applications* (IMA Volumes in Mathematics and its Applications), W. Fleming and P. I. Lions, Eds., New York: Springer-Verlag, 1988, pp. 147–186.
- [8] J. M. Harrison and L. M. Wein, "Scheduling networks of queues: Heavy traffic analysis of a two-station closed network," *Op Res*, vol. 38, no. 6, pp. 1052–1064, 1990.
- [9] M. I. Reiman, "Open queueing networks in heavy traffic," *Math Op Res*, vol. 9, pp. 441–458, 1984.
- [10] ———, "A multiclass queue in heavy traffic," *Advances in Applied Probability*, vol. 20, pp. 179–207, 1988.
- [11] W. P. Peterson, "A heavy traffic limit theorem for networks of queues with multiple customer types," *Math Op Res*, vol. 16, pp. 90–118, 1991.
- [12] J. G. Dai and Y. Wang, "Nonexistence of Brownian models of certain multiclass queueing networks," *Queueing Systems: Theory and Applications: Special Issue on Queueing Networks*, vol. 13, pp. 41–46, May 1993.
- [13] W. Whitt, "An interesting example of a multiclass queueing network," AT&T Bell Laboratories, Murray Hill, NJ, Tech. Rep., 1992.
- [14] J. G. Dai and V. Nguyen, "On the convergence of multiclass queueing networks in heavy traffic," Georgia Inst. Tech. and Massachusetts Inst. Tech., Tech. Rep., 1993.
- [15] J. F. C. Kingman, "The ergodic behaviour of random walks," *Biometrika*, vol. 48, pp. 391–396, 1961.
- [16] G. Fayolle, "On random walks arising in queueing systems: Ergodicity and transience via quadratic forms as Lyapunov functions—part I," *Queueing Systems*, vol. 5, pp. 167–184, 1989.
- [17] G. Fayolle, V. A. Malyshev, M. V. Mensikov, and A. F. Sidorenko, "Lyapunov functions for Jackson networks," INRIA, Rocquencourt, France, *Rapports de Recherche* 1380, 1991.
- [18] S. P. Meyn and D. Down, "Stability of generalized Jackson networks," *Annals Applied Probab.*, to appear, 1993.
- [19] J. R. Jackson, "Jobshop-like queueing systems," *Management Sci.*, vol. 10, pp. 131–142, 1963.
- [20] E. G. Coffman, Jr., D. S. Johnson, P. W. Shor, and R. R. Weber, "Markov chains, computer proofs and average-case analysis of best fit bin packing," *STOC 93*, to appear, 1993.
- [21] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis, "Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance," *Annals Applied Probability*, vol. 4, pp. 43–75, 1994.
- [22] S. Kumar and P. R. Kumar, "Performance bounds for queueing networks and scheduling policies," *Trans Automat Contr*, vol. 39, no. 8, pp. 1600–1611, 1994.
- [23] P. R. Kumar, "Re-entrant lines," *Queueing Systems: Theory and Applications: Special Issue on Queueing Networks*, vol. 13, pp. 87–110, May 1993.
- [24] R. W. Cottle, G. J. Habetler, and C. I. Emke, "On classes of copositive matrices," *Linear Algebra and its Applications*, vol. 3, pp. 295–310, 1970.
- [25] L. E. Andersson, G. Chang, and T. Elfving, "Criteria for copositive matrices and non negative Bezier patches," Linköping Univ. and Univ. Science and Technology, China Tech. Rep. LiTH MAT R 93/27, Aug. 1993.
- [26] K. G. Murty and S. N. Kabadi, "Some NP complete problems in quadratic and nonlinear programming," *Mathematical Programming*, vol. 39, pp. 117–129, 1987.
- [27] S. Lippman, "Applying a new device in the optimization of exponential queueing systems," *Op Res*, vol. 23, pp. 687–710, 1975.
- [28] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability* (Control and Communication in Engineering), London: Springer-Verlag, 1993.
- [29] ———, "Computable bounds for convergence rates of Markov chains," 1993.
- [30] F. P. Kelly, *Reversibility and Stochastic Networks*, New York: Wiley, 1979.



P. R. Kumar (S'77–M'77–SM'86–F'88) was born on April 21, 1952 in India. He received the D.Sc. degree from Washington University in 1977.

From 1977–1984 he was a faculty at the University of Maryland. Since 1985 he has been at the University of Illinois, Urbana-Champaign. His current research interests are in manufacturing systems, adaptive control, and learning.

Dr. Kumar is an Associate Editor at Large for *IEEE TRANSACTIONS ON AUTOMATIC CONTROL* and an Associate Editor for the following journals: *Journal of Discrete Event Dynamic Systems: Theory and Applications*, *International Journal on Adaptive Control and Signal Processing*, and *Mathematics of Control, Signals, and Systems*.



Sean P. Meyn (M'86) received the B.S. degree in mathematics from the University of California, Los Angeles, in 1982 and the M.Eng. and Ph.D. degrees in electrical engineering from McGill University, Montreal, Canada, in 1987.

He spent two years as a postdoctoral fellow at the Department of Systems Engineering, Australian National University. Since 1989 he has been an Assistant Professor with the Coordinated Science Laboratory and the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign.

Dr. Meyn serves as an Associate Editor of several journals in the systems and control area. He was a recipient of the NSF Research Initiation Award and is a member of Phi Beta Kappa. He is coauthor, with R. L. Tweedie, of the award-winning book, *Markov Chains and Stochastic Stability* (Springer-Verlag, London, 1993).

Second-Order Properties of Families of Discrete-Event Systems

Rajandran Rajan and Rajeev Agrawal, *Member, IEEE*

Abstract— We consider discrete-event systems (DES) whose logical component is characterized by a constraint set and whose temporal mechanism involves synchronization of the clock sequence with a master clock. We are interested in determining sufficient conditions on the constraint sets of a family of such synchronous DES that ensure that the event counting process of one system dominates the convex combination of the event counting processes of a collection of systems. Our point of departure is a result due to Glasserman and Yao [16], which established a sufficient condition based on characteristic functions. First we show that the characteristic function condition is equivalent to a simpler condition on the score spaces themselves. As both of these (equivalent) conditions are rather strong, however, we introduce coequality to obtain weaker sufficient conditions. To demonstrate the scope of these two results, we prove the near-concavity of the throughput in various parameters for min-linearly constrained DES. This not only covers various known concavity results for tandems, cycles, and fork-join networks of stations with general blocking and starvation, but also establishes new ones for certain classes of networks which involve splitting and merging traffic streams. These results are finally extended to the class of generalized min-linearly constrained DES.

I. INTRODUCTION

A timed discrete-event system (DES) has two aspects — the logical and the temporal. The logical aspect of the DES deals with the order in which events can occur. This is completely specified by a language which is the collection of all allowable strings of events. The temporal aspect of the DES determines which one of the allowable strings is actually realized, and the times of occurrence of the events in that string. Several mechanisms, such as state machines [1], Petri nets [2], and finitely recursive processes [3], have been proposed for generating different classes of DES languages. In this paper, we consider languages that are generated by constraint sets, $\mathcal{T} \subset \mathbb{Z}_+^m$, in a manner to be made precise in Section II. In that section, we also specify the temporal mechanism through the event counting process $\{D(t) = (D_1(t), \dots, D_m(t))\}_{t=0}^\infty$ where $D_\alpha(t)$ counts the number of occurrences of the event α up to time t . We are interested in structural properties of the event counting process $D^\mathcal{T}$ as a function of the constraint set \mathcal{T} . The main result that we obtain establishes sufficient conditions for $D^\mathcal{T}$ to be concave as a function of \mathcal{T} , in the following sense: If \mathcal{T}^p , $0 \leq p \leq r$,

satisfy $\mathcal{T}^0 \supset \left[\sum_{p=1}^r q_p \mathcal{T}^p \right]$, where

$$\left[\sum_{p=1}^r q_p \mathcal{T}^p \right] := \left\{ \left[\sum_{p=1}^r q_p \mathbf{x}^p \right] : \mathbf{x}^p \in \mathcal{T}^p, 1 \leq p \leq r \right\} \quad (1.1)$$

for some $q_p \geq 0$, $p = 1, \dots, r$, and $\sum_{p=1}^r q_p = 1$, then the corresponding event counting processes $D^p = D^{\mathcal{T}^p}$, $0 \leq p \leq r$, satisfy

$$D^0(t) \geq \sum_{p=1}^r q_p D^p(t), \quad \text{for all times } t \geq 0. \quad (1.2)$$

While this definition of concavity is not standard, it serves to ease the exposition.

The above structural result may be used to deduce the concavity or near-concavity of the throughput of certain queueing networks in various parameters such as the buffer configuration, initial job configuration, and blocking parameters. This has important implications for optimal design and control problems arising in such networks. In particular, it allows one to establish the convergence of various "local search" based optimization algorithms to the globally optimal solution. Further, when combined with other structural properties in specific queueing networks, it enables the identification of the optimal solution or the reduction of the search space for optimal allocations (see [4]–[11]). For example, consider the problem of selecting the optimal job population in a closed cycle of $M/M/1/k$ stations with communication blocking. For this problem, concavity (see [6]) along with a certain symmetry property (see [12]) implies that the throughput is maximized when the total job population is equal to half the total buffer capacity in this cycle (see [13] for some generalizations of this result). Example 5.4 in Section V-B of this paper provides another instance of how concavity may be combined with a symmetry property to solve for the optimal job configuration. For an application to optimal control, see [14, Example 6.3a], where a generalization of the dominance result (1.2) is used to extend a result on optimal routing due to Ephremides *et al.* [15].

In the next section, we develop the terminology and notation that we will require for the rest of the paper. We present a self-contained introduction to synchronous discrete-event systems (SDES), with an emphasis on treating these systems through a language/score space approach, as opposed to one that emphasizes the state evolution mechanism.

Manuscript received July 2, 1993; revised March 30, 1994. Recommended by Associate Editor, S. LaFortune. This work was supported in part by NSF Grant ECS-8919818.

The authors are with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706-1691 USA.

IEEE Log Number 9407566.

Our point of departure in Section III is a result due to Glasserman and Yao [16, Theorem 7.2], a generalization of which is presented here as Theorem 3.2. This theorem establishes that the dominance of the characteristic function of one SDES over the convex combination of the characteristic functions of a family of others implies a similar relationship among their event counting processes. The first general theoretical results in this paper are Theorem 3.3 and Lemma 3.5, which establish that dominance of one characteristic function over a convex combination of others is equivalent to a similar relation among the constraint sets of the SDES, and hence the latter too implies the dominance of one event counting process over a convex combination of others. The chief merit of this theorem is that the dominance condition for constraint sets (1.1) is much easier to verify than the dominance condition for characteristic functions.

Theorems 3.2 and 3.3 require unnecessarily strong conditions on the score space for the desired properties between event counting processes to hold. We redress this by introducing the notion of coequality in Section IV. Theorem 4.1 uses coequality to give a weaker set of conditions on the score spaces, that are still sufficient to establish the concavity of the event counting processes.

Section V-A introduces the class of min-linearly constrained SDES that model a wide variety of queueing networks. Such systems correspond to forward conflict-free Petri nets, i.e., Petri nets in which each place is followed by at most one transition (see [17]). We apply Theorem 3.3 to show the near-concavity of the event counting processes as a function of the parameter sets that define these systems (Theorem 5.3). In Section V-B we extend the class of min-linearly constrained SDES to that of generalized min-linearly constrained SDES. Such systems arise from Petri nets (not necessarily forward conflict-free) where conflict resolution is managed by a switching mechanism (see [18]). Subsequently, we generalize Theorem 5.3 to obtain near-concavity in both the parameter sets and the switching mechanism. Many existing results on the concavity of the throughput of queueing networks in various network parameters, such as the configurations of buffers and jobs, fall within the framework of min-linearly constrained SDES. In particular, these include cyclic, tandem, and fork-join networks of queues with general blocking and starvation. In addition, the min-linearly constrained SDES framework also allows us to obtain similar results for a much wider class of queueing networks which involve splitting and merging of traffic streams. We also present several examples that illustrate the scope of these models and of the earlier theoretical results.

II. THE SET UP

In this section we describe the logical and temporal aspects of SDES. The first part of this section develops the terminology of languages and score spaces, while the second defines the associated event counting process and sequences of event occurrence times. The setup as described in this section is essentially the same as that of Glasserman and Yao [16], with the difference that we focus directly on languages rather than on an underlying state evolution mechanism such as generalized semi-Markov schemes.

A. Languages and Score Spaces

The logical aspect of a discrete-event system $\Delta(\mathcal{L})$ is described by its language \mathcal{L} which is the set of all feasible sequences of events. The alphabet or event set $\mathcal{A} := \{1, 2, \dots, m\}$ is the list of all possible events that can occur in the discrete-event system Δ . A string σ of events from \mathcal{A} is given by $\sigma = \alpha_1 \cdots \alpha_n$, $\alpha_k \in \mathcal{A}$, $1 \leq k \leq n$, $n \geq 0$. When $n = 0$, we have the null string which we denote by ϕ . The universal language \mathcal{A}^* is the collection all possible strings from \mathcal{A} , i.e., $\mathcal{A}^* := \{\sigma = \alpha_1 \alpha_2 \cdots \alpha_n : \alpha_k \in \mathcal{A}, 1 \leq k \leq n, n \geq 0\}$. Any subset $\mathcal{L} \subset \mathcal{A}^*$, with $\phi \in \mathcal{L}$, is called a language. Given two string $\sigma, \rho \in \mathcal{A}^*$, we say that ρ is a prefix of σ , denoted by $\sigma \geq \rho$, if there is another string μ such that $\sigma = \rho\mu$ (μ concatenated to ρ). The score of a string σ , denoted by $[\sigma]$, is the vector $\mathbf{x} \in \mathbb{Z}_+^m$ whose i th component x_i counts the number of occurrences of i in σ , $1 \leq i \leq m$. The event $i \in \mathcal{A}$ is enabled for the string $\sigma \in \mathcal{L}$, iff $\sigma i \in \mathcal{L}$, i.e., the event i can occur after the string σ has occurred. Let the event list for the string $\sigma \in \mathcal{L}$, $\mathcal{E}(\sigma)$, be the set of events enabled for the string σ .

Definition 2.1: A language \mathcal{L} is said to be

- 1) prefix-closed iff $\sigma \in \mathcal{L}, \rho \leq \sigma \Rightarrow \rho \in \mathcal{L}$.
- 2) permutable iff $\sigma, \rho \in \mathcal{L}, [\sigma] = [\rho] \Rightarrow \mathcal{E}(\sigma) = \mathcal{E}(\rho)$.
- 3) noninterruptive iff $\alpha \in \mathcal{E}(\sigma), \sigma\mu \in \mathcal{L}, [\mu]_\alpha = 0 \Rightarrow \alpha \in \mathcal{E}(\sigma\mu)$. (Here, $[\mu]_\alpha$ denotes the α th component of $[\mu]$.)
- 4) an antimatroid with repetition iff it is prefix-closed, permutable, and noninterruptive.

Let Λ denote the set of all antimatroids with repetition, i.e., $\Lambda := \{\mathcal{L} \subset \mathcal{A}^* : \mathcal{L} \text{ is prefix-closed, permutable, and noninterruptive}\}$ (see [19] or [16]).

The subset $\mathcal{S}(\mathcal{L})$ of \mathbb{Z}_+^m given by $\mathcal{S} = \{[\sigma] : \sigma \in \mathcal{L}\}$ is called the score space of \mathcal{L} . As $\phi \in \mathcal{L}$, we always have $\mathbf{0} \in \mathcal{S}(\mathcal{L})$. Given any subset \mathcal{T} of \mathbb{Z}_+^m , with $\mathbf{0} \in \mathcal{T}$, we can recursively define a language $\mathcal{M}(\mathcal{T})$, as follows:

- $\phi \in \mathcal{M}$.
- $\sigma \in \mathcal{M}$ and $[\sigma] + \mathbf{e}^\alpha \in \mathcal{T} \Rightarrow \sigma\alpha \in \mathcal{M}$.

Here, \mathbf{e}^α is the unit vector in the α th direction. \mathcal{T} is called the constraint set for the language $\mathcal{M}(\mathcal{T})$. This is equivalent to saying that a string σ belongs to \mathcal{M} , iff the scores of all its prefixes lie in \mathcal{T} .

Definition 2.2: A subset \mathcal{T} of \mathbb{Z}_+^m is said to be

- 1) *max-closed* iff $\mathbf{x}, \mathbf{y} \in \mathcal{T}$ and $\mathbf{z} = \mathbf{x} \vee \mathbf{y} := (x_1 \vee y_1, \dots, x_m \vee y_m)$ then $\mathbf{z} \in \mathcal{T}$.
- 2) *accessible* iff for each $\mathbf{x} \in \mathcal{T}$, $\mathbf{x} \neq \mathbf{0}$, there is some $i = i(\mathbf{x})$, $1 \leq i \leq m$, such that $(\mathbf{x} - \mathbf{e}^i) \in \mathcal{T}$.

Let $\Theta := \{\mathcal{T} \subset \mathbb{Z}_+^m : \mathcal{T} \text{ is accessible and max-closed}\}$.

Given a set $\mathcal{T} \subset \mathbb{Z}_+^m$, we can construct a language $\mathcal{M}(\mathcal{T}) \subset \mathcal{A}^*$, and its score space $\mathcal{S}(\mathcal{M}(\mathcal{T})) \subset \mathbb{Z}_+^m$. Similarly, starting with a language $\mathcal{L} \subset \mathcal{A}^*$ we can construct the score space $\mathcal{S}(\mathcal{L}) \subset \mathbb{Z}_+^m$, and from that another language $\mathcal{M}(\mathcal{S}(\mathcal{L})) \subset \mathcal{A}^*$. Next, we examine the relationship between the various sets and languages constructed through the above operations.

Lemma 2.3: 1) If $\mathcal{T} \in \Theta$, then $\mathcal{M}(\mathcal{T}) \in \Lambda$ and $\mathcal{S}(\mathcal{M}(\mathcal{T})) = \mathcal{T}$.

2) If $\mathcal{L} \in \Lambda$, then $\mathcal{S}(\mathcal{L}) \in \Theta$ and $\mathcal{M}(\mathcal{S}(\mathcal{L})) = \mathcal{L}$.

Thus, there is a one to one correspondence between Θ and Λ .

Proof: 1) That $\mathcal{M}(\mathcal{T})$ is prefix-closed and permutable follows from its construction. Its noninterruptiveness can be easily deduced from the max-closure of \mathcal{T} . Again, by construction, $\mathcal{S}(\mathcal{M}(\mathcal{T})) \subset \mathcal{T}$. Conversely, observe that for any point $\mathbf{x} \in \mathcal{T}$ that is accessible from the origin, there is at least one string $\sigma \in \mathcal{M}(\mathcal{T})$ such that $[\sigma] = \mathbf{x}$. Therefore $\mathcal{S}(\mathcal{M}(\mathcal{T})) \supset \mathcal{T}$. 2) As \mathcal{L} is prefix-closed, $\mathcal{S}(\mathcal{L})$ is accessible. The max-closure of $\mathcal{S}(\mathcal{L})$ follows from the permutability and noninterruptiveness of \mathcal{L} , as shown in [16, Section 2.3]. Alternatively, let $\mathbf{x} \in \mathbb{Z}_+^m$ be a minimal vector and $\sigma, \nu \in \mathcal{L}$ be two minimal strings, such that $\mathbf{x} = [\sigma] \vee [\nu]$, but there exists no string μ such that $\sigma\mu \in \mathcal{L}$ and $\mathbf{x} = [\sigma\mu]$. Without loss of generality (WLOG), assume that $\sigma = \alpha_1 \cdots \alpha_p \neq \phi$. Note that the minimality of σ forces

$$> [\nu]. \quad (2.1)$$

Let $\sigma' := \alpha_1 \cdots \alpha_{p-1}$. By the minimality of \mathbf{x} there exists a string μ' such that $\sigma'\mu' \in \mathcal{L}$ and $[\sigma'\mu'] = [\sigma'] \vee [\nu] = \mathbf{x} - \mathbf{e}^{\alpha_p}$. Note that (2.1) implies that $[\mu']_{\alpha_p} = 0$. By the noninterruptiveness of \mathcal{L} , $\sigma'\alpha_p \in \mathcal{L} \Rightarrow \sigma'\mu'\alpha_p \in \mathcal{L}$. Next, compare σ and $\sigma'\mu'$, and apply (2.1) and noninterruptiveness to deduce that $\sigma\mu' = \sigma'\alpha_p\mu' \in \mathcal{L}$. Finally, conclude that $[\sigma\mu'] = \mathbf{x}$, which contradicts the earlier assumption. Consequently, $\mathcal{S}(\mathcal{L})$ is max-closed.

Observe that as $\mathcal{L} \in \Lambda$, a string $\sigma \in \mathcal{L}$ iff the score of all its prefixes lie in $\mathcal{S}(\mathcal{L})$. It is easy to verify this of $\mathcal{M}(\mathcal{S}(\mathcal{L}))$ too, from its construction. Hence $\mathcal{L} = \mathcal{M}(\mathcal{S}(\mathcal{L}))$. \square

Remark 2.4: In fact, we can find a set Λ' of prefix-closed and permutable languages satisfying a weaker condition than noninterruption such that there is a one-to-one correspondence between this set and the set Θ' of all accessible subsets of \mathbb{Z}_+^m . We shall not examine this more general relation further, as the main concavity results hold only for the classes Λ and Θ .

Lemma 2.5: For any $\mathcal{T} \subset \mathbb{Z}_+^m$, let \mathcal{S} be its maximal accessible subset. Then,

1) $\mathcal{M}(\mathcal{T}) = \mathcal{M}(\mathcal{S})$, and consequently $\mathcal{S}(\mathcal{M}(\mathcal{T})) = \mathcal{S}(\mathcal{M}(\mathcal{S})) = \mathcal{S}$.

2) If \mathcal{T} is max-closed then so is \mathcal{S} .

Proof: Assume WLOG that $\mathbf{0} \in \mathcal{T}$, as \mathcal{S} is empty otherwise. 1) $\mathcal{S} \subset \mathcal{T}$ implies that $\mathcal{M}(\mathcal{S}) \subset \mathcal{M}(\mathcal{T})$. By construction, the scores of all prefixes of any string in $\mathcal{M}(\mathcal{T})$ lie in \mathcal{S} . Consequently, $\mathcal{M}(\mathcal{S}) \supset \mathcal{M}(\mathcal{T})$. 2) Let $\mathcal{U} = \{\mathbf{x} \in \mathbb{Z}_+^m : \mathbf{x} = \mathbf{y} \vee \mathbf{z}; \mathbf{y}, \mathbf{z} \in \mathcal{S}, \mathbf{x} \notin \mathcal{S}\}$. Let \mathbf{u} be some minimal element¹ of \mathcal{U} . Let (\mathbf{v}, \mathbf{w}) be a minimal ordered pair such that $\mathbf{v}, \mathbf{w} \in \mathcal{S}$ and $\mathbf{u} = \mathbf{v} \vee \mathbf{w}$. Automatically, $\mathbf{u} \in \mathcal{T}$. By the accessibility of \mathcal{S} , there is some i , $1 \leq i \leq m$, such that $\hat{\mathbf{v}} = \mathbf{v} - \mathbf{e}^i \in \mathcal{S}$. By the minimality of (\mathbf{v}, \mathbf{w}) and \mathbf{u} , if we define $\hat{\mathbf{u}} := \hat{\mathbf{v}} \vee \mathbf{w}$, then $\hat{\mathbf{u}} = (\mathbf{u} - \mathbf{e}^i) \in \mathcal{S}$. This implies that \mathbf{u} is in the maximal accessible component \mathcal{S} , a contradiction that establishes that \mathcal{U} is empty. \square

Assumption 2.6: In this paper, we shall focus exclusively on DES $\Delta(\mathcal{L})$ whose language \mathcal{L} is generated from a max-closed constraint set $\mathcal{T} \subset \mathbb{Z}_+^m$, i.e., $\mathcal{L} = \mathcal{M}(\mathcal{T})$. In view of Lemmas 2.3 and 2.5, it follows that $\mathcal{L} \in \Lambda$ and that $\mathcal{S}(\mathcal{L})$ is the maximal accessible subset of \mathcal{T} .

B. Synchronous Discrete-Event Systems

A SDES, $\Delta(\mathcal{L})$, specified by a language \mathcal{L} , is a mapping that takes a sequence of clock times $\omega = (\{\omega_j^i\}_{j=1}^\infty)_{i=1}^m$ to an event counting process $\{D(t, \omega) = (D_1(t, \omega), \dots, D_m(t, \omega)) : t \geq 0\}$. We sometimes suppress dependence on ω for ease of notation. The precise mechanism by which D can be obtained from ω is described below. First, let $\{D_i^*(t) : t \geq 0\}$ be the master counting process for event i , defined by

$$T_i^*(t) := \sum_{k=1}^{\infty} \omega_i^k, \quad n \geq 0, \\ D_i^*(t) := \max\{n \geq 0 : T_i^*(n) \leq t\}, \quad t \geq 0. \quad (2.2)$$

Take $t_0 = 0$. Let t_n be the n th jump time and α_n^* the corresponding event type (direction of jump) of the process $\{D^*(t) = (D_1^*(t), \dots, D_m^*(t)) : t \geq 0\}$. Call $\sigma^{*,n} = \alpha_1^* \alpha_2^* \cdots \alpha_n^*$ the master string at time t_n induced by ω . The string $\sigma^n \in \mathcal{L}$, defined inductively below, is said to be the realized string in the SDES Δ at t_n , $n = 0, 1, \dots$

$$\sigma^0 = \phi \\ \sigma^{n+1} = \begin{cases} \sigma^n, & \text{if } \alpha_{n+1}^* \notin \mathcal{E}(\sigma^n), \\ \sigma^n \alpha_{n+1}^*, & \text{if } \alpha_{n+1}^* \in \mathcal{E}(\sigma^n). \end{cases}$$

The event counting process $D(t)$ is given by

$$D(t) = [\sigma^n] \quad \text{for } t_n \leq t < t_{n+1}.$$

Note that if $\mathcal{L} = \mathcal{A}^*$, then $\sigma^n = \sigma^{*,n}$ and $D(t) = D^*(t)$. The synchronization mechanism described here for obtaining the process $\{D(t) : t \geq 0\}$ is identical to the one described by Glasserman and Yao in [16].

Let $\{\Delta^p = \Delta(\mathcal{L}^p) : p = 0, 1, \dots, r\}$, $r \geq 1$, be a family of SDES defined on a common alphabet \mathcal{A} . Let M^r be the unit simplex in \mathbb{R}^r , i.e.,

$$M^r := \{(q_1, \dots, q_r) \in \mathbb{R}^r : \sum_{p=1}^r q_p = 1, q_p > 0, p = 1, \dots, r\}. \quad (2.3)$$

Let $\mathbf{q} \in M^r$. In this paper, we investigate sufficient conditions for the event counting process of one system to dominate the \mathbf{q} -convex combination of the event counting processes of several others, i.e., for the relation

$$D^0(t, \omega) \geq \sum_{p=1}^r q_p D^p(t, \omega), \quad \forall \text{ clock sequences } \omega, \\ \text{and times } t \geq 0 \quad (2.4)$$

to hold, in terms of conditions on the corresponding score spaces, $\{\mathcal{S}^p = \mathcal{S}(\mathcal{L}^p) : p = 0, 1, \dots, r\}$.

III. SCORE SPACE CONCAVITY

Our point of departure is a result by Glasserman and Yao [16] which establishes that if the characteristic function (defined below) of one system dominates the ‘‘average’’ of the characteristic functions of two other systems, then the event counting processes of the corresponding systems satisfy the same relation (2.4). Below, we restate a generalized version

¹ All operators on ordered tuples are taken componentwise.

of their theorem. Even though the proof in [16] continues to hold for the following theorem with only minor modifications, we reproduce it here for the sake of completeness.

Definition 3.1: Given a language $\mathcal{L} \in \Lambda$, define its characteristic function $\chi = (\chi_1, \dots, \chi_m) : \mathcal{S}(\mathcal{L}) \rightarrow \mathbb{Z}_+^m$ by $\chi_i(\mathbf{x}) = x_i + 1_{\{i \in \mathcal{E}(\mathbf{x})\}}$, $1 \leq i \leq m$, where 1_A is the indicator function of the set A . Note that if \mathcal{L} has been derived from a set $\mathcal{T} \subset \mathbb{Z}_+^m$, then $\chi(\mathbf{x}) = \max\{\mathbf{x} + \mathbf{e}^p \in \mathcal{T} : 1 \leq p \leq m\}$, where the maximum is taken componentwise, as usual.

Theorem 3.2: Let $\{\Delta^p : 0 \leq p \leq r\}$, be a family of SDES with corresponding score spaces $\{S^p : 0 \leq p \leq r\}$, characteristic functions $\{\chi^p : 0 \leq p \leq r\}$, and event counting processes $\{D^p : 0 \leq p \leq r\}$. Let $\mathbf{q} \in M^r$ be such that

$$\begin{aligned} \mathbf{x}^0 &\geq \sum_{p=1}^r q_p \mathbf{x}^p, \mathbf{x}^p \in S^p, p = 0, 1, \dots, r \\ &\Rightarrow \chi^0(\mathbf{x}^0) \geq \sum_{p=1}^r q_p \chi^p(\mathbf{x}^p). \end{aligned} \quad (3.1)$$

Then

$$D^0(t, \omega) \geq \sum_{p=1}^r q_p D^p(t, \omega) \quad \forall \text{ clock sequences } \omega, \text{ and times } t \geq 0. \quad (3.2)$$

Proof: The proof is by an induction on the successive jump times $\{t_n\}_{n=1}^\infty$ of the master counting process $\{D^*(t) : t \geq 0\}$ (see discussion around (2.2) for notation used in this proof). The induction hypothesis is that (3.2) holds for t_j , $0 \leq j \leq k$. This is trivially true for $k = 0$. Next, let $l := \alpha_k^*$ be the event that occurs in the master counting process at t_{k+1} . Note that

$$D_i^p(t_{k+1}) = \begin{cases} D_i^p(t_k), & \text{for } i \neq l, \\ \chi_i^p(D^p(t_k)) & \text{for } i = l. \end{cases}$$

In either event the induction step follows easily. \square

We now show that the dominance of one characteristic function over a convex combination of others is equivalent to a similar relationship among the corresponding score spaces as subsets of \mathbb{Z}_+^m . This makes the above result more transparent and easier to apply, as we shall show in Section V.

For any $\mathbf{q} \in M^r$ we denote

$$\sum_{p=1}^r q_p S^p := \left\{ \left[\sum_{p=1}^r q_p \mathbf{x}^p \right] : \mathbf{x}^p \in S^p, \text{ for } 1 \leq p \leq r \right\}$$

where $\lceil \mathbf{x} \rceil$ denotes the ceiling function applied componentwise, for $\mathbf{x} \in \mathbb{R}_+^m$.

Theorem 3.3: Let $\{\Delta^p : 0 \leq p \leq r\}$, be a family of SDES with corresponding score spaces $\{S^p : 0 \leq p \leq r\}$ and characteristic functions $\{\chi^p : 0 \leq p \leq r\}$. Let $\mathbf{q} \in M^r$. Consider the following assertions:

- 1) $\left[\sum_{p=1}^r q_p S^p \right] \subset S^0$, i.e., $\mathbf{x}^0 = \left[\sum_{p=1}^r q_p \mathbf{x}^p \right]$, $\mathbf{x}^p \in S^p, p = 1, \dots, r \Rightarrow \mathbf{x}^0 \in S^0$.
- 2) $\mathbf{x}^0 = \left[\sum_{p=1}^r q_p \mathbf{x}^p \right]$, $\mathbf{x}^p \in S^p, p = 0, 1, \dots, r \Rightarrow \chi^0(\mathbf{x}^0) \geq \sum_{p=1}^r q_p \chi^p(\mathbf{x}^p)$.

$$3) \mathbf{y}^0 \geq \mathbf{x}^0 = \left[\sum_{p=1}^r q_p \mathbf{x}^p \right], \mathbf{y}^0 \in S^0, \mathbf{x}^p \in S^p, p = 1, \dots, r \Rightarrow \chi^0(\mathbf{y}^0) \geq \sum_{p=1}^r q_p \chi^p(\mathbf{x}^p).$$

Then, (1) \Leftrightarrow (2) \Leftrightarrow (3).

Proof: (1) \Rightarrow (2) Fix i , $1 \leq i \leq m$. For $1 \leq p \leq r$, define $\dot{\mathbf{x}}^p \in \mathbb{Z}_+^m$ by

$$\dot{x}_j^p := \begin{cases} \chi_j^p(\mathbf{x}^p), & \text{if } j = i, \\ x_j^p, & \text{otherwise.} \end{cases}$$

Note that $\dot{\mathbf{x}}^p \in S^p$ for $1 \leq p \leq r$, by definition of the characteristic function. Next by Assertion (1), $\mathbf{x}^0 \in S^0$ and

$$\dot{\mathbf{x}}^0 := \left[\sum_{p=1}^r q_p \dot{\mathbf{x}}^p \right] \in S^0.$$

Clearly, $\mathbf{x}^0 \leq \dot{\mathbf{x}}^0 \leq \mathbf{x}^0 + \mathbf{e}^i$. From the definition of the characteristic function, $\chi^0(\mathbf{x}^0) \geq \dot{x}_i^0$. Therefore

$$(\chi_i^0(\mathbf{x}^0)) \geq \dot{x}_i^0 = \left[\sum_{p=1}^r q_p \chi_i^p(\mathbf{x}^p) \right]. \quad 1 \leq i \leq m.$$

Since i was arbitrary, it follows that $\chi^0(\mathbf{x}^0) \geq \sum_{p=1}^r q_p \chi^p(\mathbf{x}^p)$, and consequently Assertion (2) holds.

(2) \Rightarrow (1) Assume that $\left[\sum_{p=1}^r q_p S^p \right] \setminus S^0$ is nonempty and let \mathbf{x} be some minimal element of this set. Let $(\sigma^1, \sigma^2, \dots, \sigma^r) \in \mathcal{L}^1 \times \mathcal{L}^2 \times \dots \times \mathcal{L}^r$ be a minimal ordered r -tuple of strings (in the prefix-ordering taken componentwise), such that

$$\mathbf{x} = \left[\sum_{p=1}^r q_p [\sigma^p] \right].$$

Let us write σ^p as $i_1^p i_2^p \dots i_{s_p}^p$, where $i_j^p \in \mathcal{A}$, $1 \leq j \leq s_p$, $1 \leq p \leq r$. We allow $s_p = 0$ if $\sigma^p = \phi$. As $0 \in S^0$, WLOG, we may take $q_1 > 0$, $s_1 \geq 1$ and $i_1^1 = 1$. Define $\rho^1 = i_1^1 i_2^1 \dots i_{s_1-1}^1$, and let

$$\mathbf{y} := \left[(q_1 [\rho^1]) + \sum_{p=2}^{r-1} q_p [\sigma^p] \right].$$

By minimality of \mathbf{x} and $(\sigma^1, \sigma^2, \dots, \sigma^r)$, we have $\mathbf{y} = \mathbf{x} - \mathbf{e}^1$ and $\mathbf{y} \in S^0$. Since $\mathbf{x} \notin S^0$, it follows that $\chi_1^0(\mathbf{y}) = y_1$. Moreover, $\chi_1^1([\rho^1]) = [\rho^1]_1 + 1 = [\sigma^1]_1$. From this we have

$$\begin{aligned} q_1 \chi_1^1([\rho^1]) + \sum_{p=2}^r q_p \chi_1^p([\sigma^p]) &\geq \sum_{p=1}^r q_p [\sigma^p]_1 \\ &= x_1 \\ &> y_1 \\ &= \chi_1^0(\mathbf{y}) \end{aligned}$$

which contradicts (2). Hence the result.

(1) & (2) \Rightarrow (3) This follows directly from the noninterruptiveness condition, which implies the monotonicity of the characteristic function, i.e.,

$$\mathbf{y}^0, \mathbf{x}^0 \in S^0, \mathbf{y}^0 \geq \mathbf{x}^0 \Rightarrow \chi^0(\mathbf{y}^0) \geq \chi^0(\mathbf{x}^0).$$

(See [16, Theorem 2.3] for more details).

(3) \Rightarrow (2) Obvious. \square

In light of Theorem 3.2, any of the above equivalent conditions implies (3.2).

Remark 3.4: Letting $r = 1$ in Theorem 3.3, we see that the event counting processes of SDES are monotone as functions of the score spaces. This result is the monotonicity of the event counting processes across schemes proved by Glasserman and Yao for synchronized GSMS [20, Theorem 6.4]. For the case of $r \geq 2$, the above theorem provides a score space concavity result analogous to the score space monotonicity result.

In the lemma below we show that it suffices to check Assertion 1 of the above theorem, with the constraint set \mathcal{T}^p instead of the score space S^p , where $\mathcal{T}^p \subset \mathbb{Z}_+^m$ generates the SDES Δ^p .

Lemma 3.5: Let $\{\mathcal{T}^p : 0 \leq p \leq r\}$ be subsets of \mathbb{Z}_+^m and let $\{S^p : 0 \leq p \leq r\}$ be their corresponding maximal accessible subsets. Then

$$\mathcal{T}^0 \supset \left[\sum_{p=1}^r q_p \mathcal{T}^p \right] \Rightarrow S^0 \supset \left[\sum_{p=1}^r q_p S^p \right].$$

Proof: Assume that $\mathbf{x}^p \in S^p$, $\mathbf{x}^0 = \left[\sum_{p=1}^r q_p \mathbf{x}^p \right] \in \mathcal{T}^0 \setminus S^0$, $1 \leq p \leq r$, and that $(\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^r)$ are a componentwise minimal tuple satisfying this. Without loss of generality, assume that $\mathbf{x}^1 - \mathbf{e}^1 \in S^1$. Let $\mathbf{y}^0 := \left[q_1(\mathbf{x}^1 - \mathbf{e}^1) + \sum_{p=2}^r q_p \mathbf{x}^p \right]$. By the minimality of $(\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^r)$, $\mathbf{y}^0 = \mathbf{x}^0 - \mathbf{e}^1 \in S^0$. By accessibility, $\mathbf{x}^0 \in S^0$, which contradicts our earlier assumption. Hence the result. \square

IV. COEVALITY

Theorems 3.2 and 3.3 require unnecessarily strong conditions on the score space to obtain the dominance of one event counting process over the convex combination of several others. There, we require (3.1) of Theorem 3.2 to hold for all choices of $\{\mathbf{x}^p \in S^p : 0 \leq p \leq r\}$. Note, however, that we really need to verify that condition only for those $\{\mathbf{x}^p \in S^p : 0 \leq p \leq r\}$ that can occur simultaneously. In this section we introduce the notion of coeuality to exploit this basic idea and to obtain weaker conditions for one event counting process to dominate a convex combination of others.

Let $\{\Delta^p : 0 \leq p \leq r\}$ be a family of SDES. We say that $(\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^r)$ is coeval, iff there exists a clock sequence ω and a time $t \geq 0$, such that the string σ^p realized in Δ^p at time t has score $[\sigma^p] = \mathbf{x}^p$ (See Section II-B). Then, σ^* , the master string at time t , is called an enabling string of $(\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^r)$. Let

$$S^{0,1,\dots,r} := \{(\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^r) \in S^0 \times S^1 \times \dots \times S^r : (\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^r) \text{ is coeval}\}$$

and let

$$S^{1,2,\dots,r} := \{(\mathbf{x}^1, \dots, \mathbf{x}^r) \in S^1 \times \dots \times S^r : (\mathbf{x}^1, \dots, \mathbf{x}^r) \text{ is coeval}\}.$$

Theorem 4.1: Let $\{\Delta^p : 0 \leq p \leq r\}$ be a family of SDES with corresponding score spaces $\{S^p : 0 \leq p \leq r\}$ and characteristic functions $\{\chi^p : 0 \leq p \leq r\}$. Let $q \in \mathbb{Q}$. Consider the following assertions:

- 1) $\mathbf{x}^0 = \left[\sum_{p=1}^r q_p \mathbf{x}^p \right]$, $(\mathbf{x}^1, \dots, \mathbf{x}^r) \in S^{1,2,\dots,r}$, $\mathbf{x}^0 \in S^0$.
- 2) $\mathbf{x}^0 = \left[\sum_{p=1}^r q_p \mathbf{x}^p \right]$, $\mathbf{x}^0 \in S^0$, $(\mathbf{x}^1, \dots, \mathbf{x}^r) \in S^{1,2,\dots,r} \Rightarrow \chi^0(\mathbf{x}^0) \geq \sum_{p=1}^r q_p \chi^p(\mathbf{x}^p)$.
- 3) $\mathbf{y}^0 \geq \mathbf{x}^0 = \left[\sum_{p=1}^r q_p \mathbf{x}^p \right]$, $\mathbf{y}^0 \in S^0$, $(\mathbf{x}^1, \dots, \mathbf{x}^r) \in S^{1,2,\dots,r} \Rightarrow \chi^0(\mathbf{y}^0) \geq \sum_{p=1}^r q_p \chi^p(\mathbf{x}^p)$.
- 4) $(\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^r) \in S^{0,1,2,\dots,r} \Rightarrow \mathbf{x}^0 \geq \sum_{p=1}^r q_p \mathbf{x}^p$.

Then, (1) \Leftrightarrow (2) \Leftrightarrow (3) \Rightarrow (4).

Proof: (1) \Rightarrow (2) Let $(\mathbf{x}^1, \dots, \mathbf{x}^r) \in S^{1,2,\dots,r}$. By Assertion (1), $\mathbf{x}^0 := \left[\sum_{p=1}^r q_p \mathbf{x}^p \right] \in S^0$. Fix i , $1 \leq i \leq m$. For $1 \leq p \leq r$, define $\hat{\mathbf{x}}^p$ by

$$\hat{x}_j^p := \begin{cases} x_j^p(\mathbf{x}^p), & \text{if } j = i, \\ x_j^p & \text{otherwise.} \end{cases}$$

By the definition of the characteristic functions, it is easy to see that $(\hat{\mathbf{x}}^1, \dots, \hat{\mathbf{x}}^r) \in S^{1,2,\dots,r}$. Therefore by Assertion (1)

$$\hat{\mathbf{x}}^0 := \left[\sum_{p=1}^r q_p \hat{\mathbf{x}}^p \right] \in S^0. \quad (4.1)$$

Clearly, $\mathbf{x}^0 \leq \hat{\mathbf{x}}^0 \leq \mathbf{x}^0 + \mathbf{e}^i$. Again, from the definition of the characteristic function $\chi^0(\mathbf{x}^0) \geq \chi^0(\hat{\mathbf{x}}^0)$, which implies that

$$\chi_i^0(\mathbf{x}^0) \geq \hat{x}_i^0 = \left[\sum_{p=1}^r q_p \chi_i^p(\mathbf{x}^p) \right]. \quad (4.2)$$

Since i was arbitrary, it follows that $\chi^0(\mathbf{x}^0) \geq \sum_{p=1}^r q_p \chi^p(\mathbf{x}^p)$, and consequently Assertion (2) holds.

(2) \Rightarrow (1) Let $\mathcal{U} = \{(\mathbf{x}^1, \dots, \mathbf{x}^r) \in S^{1,2,\dots,r} : \left[\sum_{p=1}^r q_p \mathbf{x}^p \right] \notin S^0\}$ be nonempty, and let $(\mathbf{x}^1, \dots, \mathbf{x}^r)$ be some (componentwise) minimal element of \mathcal{U} , and $\sigma^{*,s} = \alpha_1^* \alpha_2^* \dots \alpha_s^*$ be some (prefixwise) minimal enabling string of $(\mathbf{x}^1, \dots, \mathbf{x}^r)$. Let $\sigma^{*,s-1} = \alpha_1^* \alpha_2^* \dots \alpha_{s-1}^*$ be the enabling string of $(\hat{\mathbf{x}}^1, \dots, \hat{\mathbf{x}}^r)$, for $\hat{\mathbf{x}}^p \in S^p$.

First, note that $(\hat{\mathbf{x}}^1, \dots, \hat{\mathbf{x}}^r) \in S^{1,2,\dots,r}$, and $\chi_{\alpha_s^*}^p(\hat{\mathbf{x}}^p) = x_{\alpha_s^*}^p$, for all p . Further, the minimality of $(\mathbf{x}^1, \dots, \mathbf{x}^r) \in \mathcal{U}$ and $\sigma^{*,s}$ implies that

$$\hat{\mathbf{x}}^0 := \left[\sum_{p=1}^r q_p \hat{\mathbf{x}}^p \right] = \left[\sum_{p=1}^r q_p \mathbf{x}^p \right] - \mathbf{e}^{\alpha_s^*} \quad (4.3)$$

and that $\hat{\mathbf{x}}^0 \in S^0$. Using the fact that $\hat{\mathbf{x}}^0 + \mathbf{e}^{\alpha_s^*} \notin S^0$ we conclude that

$$\begin{aligned} \chi_{\alpha_s^*}^0(\hat{\mathbf{x}}^0) &= \hat{x}_{\alpha_s^*}^0 \\ &< \left[\sum_{p=1}^r q_p x_{\alpha_s^*}^p \right] \quad \text{by (4.3)} \\ &= \left[\sum_{p=1}^r q_p \chi_{\alpha_s^*}^p(\mathbf{x}^p) \right] \end{aligned}$$

which contradicts Assertion (2). Therefore \mathcal{U} is empty and Assertion (1) holds.

(1)&(2) \Rightarrow (3) This follows directly from the noninterruptiveness condition.

(3) \Rightarrow (2) Obvious.

(3) \Rightarrow (4) Assertion (4) is a restatement of (3.2) of Theorem 3.2. The proof is very similar to that of Theorem 3.2. \square

The simplest situation in which coequality proves itself strictly more general than the score space result (Theorem 3.3) is when the DES Δ has a nonconvex score-space, \mathcal{S} , i.e., when the ceiling of a convex combination of scores from \mathcal{S} falls outside \mathcal{S} . This happens, for example, when $\mathcal{S} = \{x \in \mathbb{Z}_+^2 : x_2 \leq \frac{1}{2}x_1\}$. Then, $x = (0, 0)$, $y = (2, 1) \in \mathcal{S}$, but $z = \lceil (x + y)/2 \rceil = (1, 1) \notin \mathcal{S}$. Since $\mathcal{S} \not\supseteq (\mathcal{S} + \mathcal{S})/2$, it is not possible to use the score space result to prove the obvious result that $D(t) \geq \frac{1}{2}(D(t) + D(t))$. But in this case, the coeval sets are easy to characterize, and therefore we can show from Theorem 4.1 that relation (3.2) holds among the event counting processes.

It is not usually necessary to explicitly characterize coeval sets across different systems. Often, Theorem 3.3 is inapplicable because of the existence of tuples (x^1, \dots, x^r) , $x^p \in S^p$, $p = 1, \dots, r$, whose convex combination falls outside S^0 . If it can be shown that these tuples are noncoeval, then Theorem 4.1 can be used to prove that one event counting process dominates the convex combination of others.

V. APPLICATIONS TO CONSTRAINED SDES

In Section II, we explained how SDES may be generated from constraint sets $\mathcal{T} \subset \mathbb{Z}_+^m$. Typically, the constraint sets that we shall examine will have the form

$$\mathcal{T} = \{x \in \mathbb{Z}_+^m : x \leq f(x)\}$$

for some increasing constraint function, $f : \mathbb{Z}_+^m \rightarrow \mathbb{Z}_+^m$. Such constrained SDES arise naturally in many queueing networks, as may be seen from the examples to follow. In Section V-A, we introduce min-linearly constrained SDES and show that the structure of the event counting process of such systems relates naturally to the structure of their constraining family of functions. Subsequently, in Section V-B, we extend this to generalized min-linearly constrained SDES.

A. Min-Linearly Constrained SDES

Let

$$A := \{a_{ijk} : 1 \leq i, j \leq m, 1 \leq k \leq n\},$$

$$B := \{b_{jk} : 1 \leq j \leq m, 1 \leq k \leq n\}$$

be two parameter sets of nonnegative real numbers. Consider the set $\mathcal{T} \subset \mathbb{Z}_+^m$ defined by a min-linear constraint function based on the above parameter sets

$$\mathcal{T}(A, B) = \left\{ x \in \mathbb{Z}_+^m : x_j \leq \min_{1 \leq k \leq n} \left\{ \sum_{i=1}^m a_{ijk} x_i + b_{jk} \right\}, \quad 1 \leq j \leq m \right\}. \quad (5.1)$$

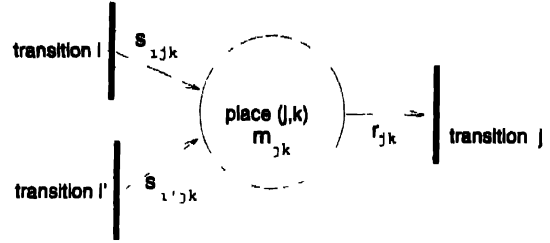


Fig. 1. Portion of a forward conflict-free Petri-net

It is easy to see that $0 \in \mathcal{T}$ and that \mathcal{T} is max-closed. Consequently, by Lemmas 2.3 and 2.5, its maximal accessible component $\mathcal{S}(A, B)$ can be used to define a prefix-closed, permutable and noninterruptive language $\mathcal{L}(A, B)$. The SDES $\Delta(A, B)$ defined by $\mathcal{L}(A, B)$ is said to be min-linearly constrained.

Remark 5.1: When $n = m$, $a_{ijk} = 1$ if $j \neq i = k$, $1 \leq i, j, k \leq m$, and $a_{ijk} = 0$ otherwise, the set $\mathcal{T} \subset \mathbb{Z}_+^m$ is given by

$$\mathcal{T}(A, B) = \left\{ x \in \mathbb{Z}_+^m : x_j \leq x_k + b_{jk}, 1 \leq j, k \leq m, j \neq k \right\}. \quad (5.2)$$

This corresponds to the class of all conflict-free Petri nets² (also called decision-free), i.e., Petri nets where each place is preceded and followed by exactly one transition. Baccelli and Liu [2] have shown the concavity of the throughput of such Petri nets in the initial marking. Several important classes of networks for which concavity results are already known, such as tandems and cycles of queues with general blocking and starvation [21], [13], and fork-join networks [21], are special cases of the above. In such networks, the constraint in (5.2) can be interpreted as restricting the number of jobs processed by server j from exceeding the number of jobs processed by any other server k , by more than a fixed quantity b_{jk} .

The class of min-linearly constrained SDES includes the much larger class of structurally forward conflict-free Petri nets, i.e., Petri nets with at most one transition following each place, but with no restrictions on the number of transitions that may precede a place (see [17]). Thus, in such a Petri-net, a place preceding a transition j cannot precede another transition $i \neq j$. WLOG, we assume that exactly n places precede each transition j , and denote each such place by the pair (j, k) , $k = 1, \dots, n$, where $j = 1, \dots, m$ are the possible transitions. Let there be m_{jk} tokens present initially in place (j, k) . We assume that the firing of transition (occurrence of event) i deposits s_{ijk} tokens in place (j, k) , and that transition j requires r_{jk} tokens from the place (j, k) to fire. Thus, when the score is $x = (x_1, \dots, x_m)$, i.e., transition i has fired x_i times, $i = 1, \dots, m$, then the transition j is enabled iff

$$(x_j + 1) \leq \min_{1 \leq k \leq n} \left\{ \left[\sum_{i=1}^m s_{ijk} x_i + m_{jk} \right] / r_{jk} \right\}.$$

It is easy to see that the constraint set of this system is precisely $\mathcal{T}(A, B)$ given by (5.1) with $a_{ijk} = s_{ijk}/r_{jk}$ and $b_{jk} = m_{jk}/r_{jk}$.

²Petri nets are a special class of discrete-event systems, with transitions in Petri nets corresponding to events in SDES.

In the next theorem, we apply Theorem 3.3 to investigate the effect of varying the parameter set B (with a fixed A) on the event counting process D of the min-linearly constrained SDES $\Delta(A, B)$. For this we need to define the slack $s(q)$ associated with any $q \in M'$ as

$$s(q) = \sup \left\{ \left[\sum_{p=1}^r q_p t^p \right] - \sum_{p=1}^r q_p t^p \mid t^p \in \mathbb{Z}_+, p=1, \dots, r \right\} \quad (5.3)$$

In the following theorem we give an explicit formula for the slack

Proposition 5.2 Let $q = (q_1, \dots, q_r) \in M'$

1) If q is rational, i.e., if $q_p = \frac{t_p}{\delta_p}$ where t_p and δ_p are co-prime integers for $p = 1, \dots, r$, then

$$s(q) = 1 - \frac{1}{\text{lcm}(\delta_1, \dots, \delta_r)}$$

Here, $\text{lcm}(\delta_1, \dots, \delta_r)$ denotes the least common multiple of $\delta_1, \dots, \delta_r$.

2) If q is irrational, i.e., q_l is irrational for some $p = 1, \dots, r$, then $s(q) = 1$.

Proof See Appendix \square

Theorem 5.3 Let $\{\Delta(A, B^p) \mid 0 \leq p \leq r\}$ be a family of min-linearly constrained SDES with corresponding score spaces $\{S^p \mid 0 \leq p \leq r\}$ and departure processes $\{D^p \mid 0 \leq p \leq r\}$. Let $q \in M'$ be such that for all $1 \leq j \leq m$, $1 \leq l \leq n$

$$b_{jk}^0 \geq \sum_{i=1}^r q_i b_{jk}^i + s(q) \quad (5.4)$$

Then $S^0 \supset \left[\sum_{i=1}^r q_i S^i \right]$ and consequently

$$D^0(t, \omega) \geq \sum_{i=1}^r q_i D^i(t, \omega) \quad \forall \text{ clock sequences } \omega \quad (5.5)$$

and times $t > 0$

In addition, if $\{a_{ijk} \mid 1 \leq i \leq m\}$ are all nonnegative integers for some (j, k) , $1 \leq j \leq m$, $1 \leq k \leq n$, then for those (j, k) (5.4) can be replaced by the weaker condition

$$b_{jk}^0 \geq \left[\sum_{i=1}^r q_i b_{jk}^i \right] \quad (5.6)$$

Proof In light of Theorem 3.3 and Lemma 3.5, it is enough to show that $T(A, B^0) \supset \left[\sum_{i=1}^r q_i T(A, B^i) \right]$. Let $y^p \in T(A, B^p)$, $p = 1, \dots, r$, and define $y^0 = \left[\sum_{i=1}^r q_i y^i \right]$. Then for any j , $1 \leq j \leq m$

$$\begin{aligned} y_j^0 &= \left[\sum_{p=1}^r q_p y_j^p \right] \\ &\leq \left[\sum_{p=1}^r q_p \min_{1 \leq k \leq n} \left[\sum_{i=1}^m a_{ijk} y_i^p + b_{jk}^p \right] \right] \\ &\leq \min_{1 \leq k \leq n} \left[\sum_{p=1}^r q_p \left[\sum_{i=1}^m a_{ijk} y_i^p + b_{jk}^p \right] \right] \quad (5.7) \end{aligned}$$

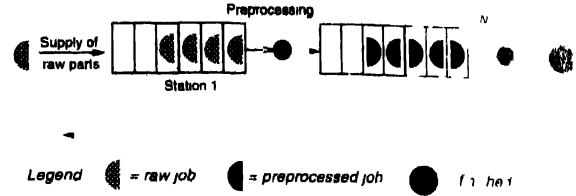


Fig. 2. Closed welding type join network.

Moreover, for any k , $1 \leq k \leq n$

$$\begin{aligned} \left[\sum_{i=1}^r q_i \left[\sum_{j=1}^m a_{ijk} y_j^i + b_{jk}^i \right] \right] &\leq \sum_{p=1}^r q_p \left[\sum_{j=1}^m a_{ijk} y_j^p + b_{jk}^p \right] + s(q) \\ &\leq \sum_{i=1}^m a_{ijk} \sum_{j=1}^r q_j y_j^i + \sum_{i=1}^m q_i b_{jk}^i + s(q) \\ &\leq \sum_{i=1}^m a_{ijk} y_i^0 + b_{jk}^0 \quad (\text{by 5.4}) \quad (5.8) \end{aligned}$$

Thus $y^0 \in T(A, B^0)$.

If $\{a_{ijk} \mid 1 \leq i \leq m\}$ are all nonnegative integers for some k , $1 \leq k \leq n$, then we have

$$\begin{aligned} \left[\sum_{i=1}^r q_i \left[\sum_{j=1}^m a_{ijk} y_j^i + b_{jk}^i \right] \right] &\leq \left[\sum_{i=1}^m a_{ijk} \sum_{j=1}^r q_j y_j^i + \sum_{i=1}^r q_i b_{jk}^i \right] \\ &\leq \sum_{i=1}^m a_{ijk} \left[\sum_{j=1}^r q_j y_j^i \right] + \left[\sum_{i=1}^r q_i b_{jk}^i \right] \\ &< \sum_{i=1}^m a_{ijk} y_i^0 + b_{jk}^0 \quad (\text{by 5.6}) \end{aligned}$$

In this case too $y^0 \in T(A, B^0)$ \square

Note that most of the concavity results obtained so far in the literature for specific SDES fall within the scope of Remark 5.1 and condition (5.6) of Theorem 5.3. We now present an example to illustrate how Theorem 5.3 can be used to obtain near concavity results for more complex systems.

Example 5.4 (Welding Station With Constant Work In Progress (WIP)) Consider the two-station cycle shown in Fig. 2. This models a production line with constant WIP where each pair of successive jobs that depart from station 1 is welded together at station 2. As soon as a job leaves station 2, two parts enter station 1 simultaneously. We assume that station i , $i = 1, 2$, has a buffer of size k_i with $0 \leq J \leq k_i$ jobs initially. Server i is blocked (communication blocking) if the buffer downstream is full. Let $D_i(t)$ denote the number of jobs that complete service at station i up to time t for $i = 1, 2$. Note that at any time t , $D_1(t)$ cannot exceed twice $D_2(t)$ by more than the number of jobs initially present at station 1 or the number of empty spaces initially present at station 2. Similarly, $D_2(t)$ cannot exceed one half $D_1(t)$ by more than the number of jobs initially present at station 2 or the number of empty spaces initially present at station 1. From this, we can see that the constraint set T that specifies the

SDES corresponding to this network is given by

$$\mathcal{T} = \left\{ \mathbf{x} \in \mathbb{Z}_+^2 : x_1 \leq 2x_2 + ((k_2 - J_2) \wedge J_1) \right. \\ \left. x_2 \leq \frac{1}{2}J_1 + \frac{1}{2}((k_1 - J_1) \wedge J_2) \right\}.$$

This is a min-linearly constrained SDES with appropriately defined parameter sets \mathbf{A} and \mathbf{B} .

Let $\{\Delta(\mathbf{A}, \mathbf{B}^p) : p = 0, 1, \dots, r\}$ be a synchronized family of such systems. Then the conditions (5.4)–(5.6) are satisfied with $\mathbf{q} \in M^r$, if

$$k_2^0 - J_2^0 \geq \left[\sum_{p=1}^r q_p (k_2^p - J_2^p) \right], \\ \frac{1}{2}(k_1^0 - J_1^0) \geq \sum_{p=1}^r q_p \frac{1}{2}(k_1^p - J_1^p) + s(\mathbf{q}), \\ J_1^0 \geq \left[\sum_{p=1}^r q_p J_1^p \right], \\ \frac{1}{2}J_2^0 \geq \sum_{p=1}^r q_p \frac{1}{2}J_2^p + s(\mathbf{q}).$$

We may deduce the near-concavity of the throughput jointly in the number of jobs and in the number of empty spaces.

Remark 5.5: The slack $s(\mathbf{q})$ obtained through Lemma 5.3 cannot be tightened in general, in the sense that there do exist families of systems where that slack is the best we can do. For instance, take $r = 2$, $J_1^0 = 1$, $J_1^1 = 0$ and $J_2^1 = 2$, and $J_2^0 = J_2^1 = J_2^2 = 0$. In this case no departures take place from station 2 of the systems Δ^0 and Δ^1 , while departures do occur in the system Δ^2 . Consequently, the $D^0(t)$ does not dominate a \mathbf{q} -convex combination of the other two, with $\mathbf{q} = (\frac{1}{2}, \frac{1}{2})$ (say). On the other hand, for certain families of min-linearly constrained SDES the slack can indeed be tightened by more careful analysis.

B. Generalized Min-Linearly Constrained SDES

Let the parameter set \mathbf{A} be as before, and let

$$\mathbf{I} = \{I_{jk} : \mathbf{R}_+ \rightarrow \mathbf{R}_+; k = 1, \dots, n, j = 1, \dots, m\}$$

be a set of nondecreasing switching functions. Consider the set $\mathcal{T} \subset \mathbb{Z}_+^m$ defined by a generalized min-linear constraint function based on the parameter \mathbf{A} and the switching function \mathbf{I} as follows

$$\mathcal{T}(\mathbf{A}, \mathbf{I}) := \left\{ \mathbf{x} \in \mathbb{Z}_+^m : x_j \leq \min_{1 \leq k \leq n} \left\{ I_{jk} \left(\sum_{i=1}^m a_{ijk} x_i \right) \right\}, \quad 1 \leq j \leq m \right\}.$$

This represents an extension of min-linearly constrained SDES model, as the constraint $\sum_{i=1}^m a_{ijk} x_i + b_{jk}$ has been replaced by $I_{jk}(\sum_{i=1}^m a_{ijk} x_i)$, where I_{jk} is a nondecreasing function. A SDES which is defined from the maximal accessible subset of the set $\mathcal{T}(\mathbf{A}, \mathbf{I})$ is said to be a generalized min-linearly constrained SDES. This allows us to model a bigger class of Petri nets — those for which structural consumption conflicts

are resolved through a predefined “switching” mechanism (see [18]).

We have the following near-concavity result for generalized min-linearly constrained SDES.

Theorem 5.6: Let $\{\Delta(\mathbf{A}, \mathbf{I}^p) : 0 \leq p \leq r\}$ be a family of generalized min-linearly constrained SDES with corresponding score spaces $\{S^p : 0 \leq p \leq r\}$, and departure processes $\{D^p : 0 \leq p \leq r\}$. Let $\mathbf{q} \in M^r$ be such that for all $\mathbf{x}^0 = \left[\sum_{p=1}^r q_p \mathbf{x}^p \right]$, $\mathbf{x}^p \in S^p$, $1 \leq p \leq r$, and for all $1 \leq j \leq m$, $1 \leq k \leq n$

$$I_{jk}^0 \left(\sum_{i=1}^m a_{ijk} x_i^0 \right) \geq \left[\sum_{p=1}^r q_p \left[I_{jk}^p \left(\sum_{i=1}^m a_{ijk} x_i^p \right) \right] \right]. \quad (5.9)$$

Then $S^0 \supset \left[\sum_{p=1}^r q_p S^p \right]$, and consequently,

$$D^0(t, \omega) \geq \sum_{p=1}^r q_p D^p(t, \omega), \quad \forall \text{ clock sequences } \omega, \\ \text{and times } t \geq 0. \quad (5.10)$$

Proof Proceeds on the same lines as Theorem 5.3. Let $\mathbf{y}^p \in \mathcal{T}(\mathbf{A}, \mathbf{I}^p)$, $p = 1, \dots, r$ and define $\mathbf{y}^0 = \left[\sum_{p=1}^r q_p \mathbf{y}^p \right]$. Then for any j , $1 \leq j \leq m$

$$y_j^0 = \left[\sum_{p=1}^r q_p y_j^p \right] \\ \leq \left[\sum_{p=1}^r q_p \left[\min_{1 \leq k \leq n} I_{jk}^p \left(\sum_{i=1}^m a_{ijk} y_i^p \right) \right] \right] \\ = \left[\sum_{p=1}^r q_p \min_{1 \leq k \leq n} \left[I_{jk}^p \left(\sum_{i=1}^m a_{ijk} y_i^p \right) \right] \right] \\ \leq \min_{1 \leq k \leq n} \left[\sum_{p=1}^r q_p \left[I_{jk}^p \left(\sum_{i=1}^m a_{ijk} y_i^p \right) \right] \right] \\ \leq \min_{1 \leq k \leq n} \left\{ I_{jk}^0 \left(\sum_{i=1}^m a_{ijk} y_i^0 \right) \right\}.$$

Thus, $\mathbf{y}^0 \in \mathcal{T}(\mathbf{A}, \mathbf{I}^0)$. \square

Note that the above result strictly generalizes Theorem 5.3. The following lemma can be used to deduce the near-concavity of Petri nets with conflict resolution in the initial markings.

Lemma 5.7: Suppose that for some j, k

$$I_{jk}^p(\cdot) = H_{jk}^p(\cdot) + b_{jk}^p, \quad p = 0, 1, \dots, r$$

for some nondecreasing function $H_{jk}^p(\cdot)$, and that there exists a constant c_{jk} such that

$$b_{jk}^0 \geq \sum_{p=1}^r q_p b_{jk}^p + s(\mathbf{q}) + c_{jk},$$

and that for all $z^0, z^1, \dots, z^r \in \mathbf{R}_+$, with $z^0 \geq \sum_{p=1}^r q_p z^p$,

$$H_{jk}^0(z^0) + c_{jk} \geq \sum_{p=1}^r q_p H_{jk}^p(z^p). \quad (5.11)$$

Then the $\{I_{jk}^p : p = 0, 1, \dots, r\}$ satisfy (5.9).

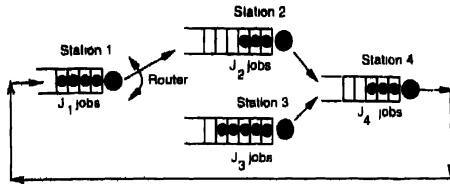


Fig. 3 Closed network with splitting and merging

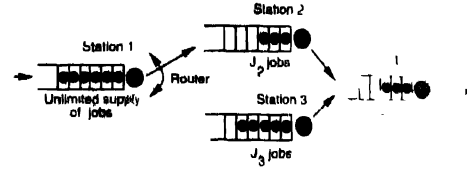


Fig. 4 Open network with splitting and merging

Proof Let $\mathbf{x}^p \in \mathcal{S}^j$, $1 \leq p \leq i$, and $\mathbf{x}^0 = M^j$ be such that $[\sum_{p=1}^i q_p \mathbf{x}^p]$. Then

$$\begin{aligned} I_{jk}^0 \left(\sum_{i=1}^m a_{ijk} x_i^0 \right) &= H_{jk}^0 \left(\sum_{i=1}^m a_{ijk} x_i^0 \right) + b_{jk}^0 \\ &\geq H_{jk}^0 \left(\sum_{i=1}^i q_i \sum_{i=1}^m a_{ijk} x_i^i \right) \\ &\quad + \sum_{i=1}^i q_i b_{jk}^i + s(\mathbf{q}) + \alpha_{jk} \\ &\geq \sum_{i=1}^i q_i H_{jk}^i \left(\sum_{i=1}^m a_{ijk} x_i^i \right) \\ &\quad + \sum_{i=1}^i q_i b_{jk}^i + s(\mathbf{q}) \\ &\geq \left[\sum_{i=1}^i q_i \left(H_{jk}^i \left(\sum_{i=1}^m a_{ijk} x_i^i \right) + b_{jk}^i \right) \right] \\ &= \left[\sum_{i=1}^i q_i \left[I_{jk}^i \left(\sum_{i=1}^m a_{ijk} x_i^i \right) \right] \right] \end{aligned}$$

□

Example 5.8 (Closed Network with Splitting and Merging)
Consider the closed network shown in Fig. 3. The i th job to depart station 1 is sent to station $H(i) \in \{2, 3\}$ where $H(\cdot)$ is some routing (splitting) mechanism. Define $H_j(i) = \sum_{q=1}^i 1_{\{H(q)=j\}}$, $j = 2, 3$. The output streams from these two stations are merged into station 4 from where they cycle back to station 1.

We assume that there is no blocking, i.e. all buffers have unlimited capacities. Let $\mathbf{J} = (J_1, J_2, J_3, J_4)$ be the initial configuration of jobs in the system. The constraint set \mathcal{T} that specifies the SDES corresponding to this network is given by

$$\begin{aligned} \mathcal{T} = \{ \mathbf{x} \in \mathbb{Z}_+^4 : & i_1 \leq i_1 + J_1 \\ & i_2 \leq H_2(i_1) + J_2 \\ & i_3 \leq H_3(i_1) + J_3 \\ & i_4 \leq i_2 + i_3 + J_4 \} \end{aligned}$$

Let the routing mechanism satisfy

$$\beta_j i - \epsilon_j \leq H_j(i) \leq \beta_j i + \alpha_j, \quad i \in \mathbb{Z}_+, j = 2, 3 \quad (5.12)$$

This is a generalized min-linearly constrained SDES with an appropriately defined parameter set \mathbf{A} and switching mechanism \mathbf{I} ($J_j(i) = H_j(i) + J_j$, $i \in \mathbb{Z}_+$, $j = 2, 3$). Let $\{\Delta(J^p) : 0 \leq p \leq i\}$ be a synchronized family of such systems, parameterized by the initial configurations. Let $\mathbf{q} \in$

$$\begin{aligned} I_j^0 &\geq \left[\sum_{p=1}^i q_p J_j^p \right], \quad j = 1, 1 \\ I_j^0 &= \sum_{p=1}^i q_p J_j^p + s(\mathbf{q}) + \epsilon_j + \alpha_j, \quad j = 2, 3 \quad (5.13) \end{aligned}$$

Then, (5.11) is satisfied with $\epsilon_j = \epsilon_j + \alpha_j$ for $j = 2, 3$ as for all $z^0 \in \mathbb{Z}_+^4$, $\mathbf{z}^i \in \mathbb{R}_+$ with $z^0 \geq \sum_{i=1}^i q_i \mathbf{z}^i$

$$\begin{aligned} H_j(z^0) + \epsilon_j + \alpha_j &\geq \beta_j z^0 + \alpha_j \\ &\geq \sum_{p=1}^i q_p (\beta_j z^p + \alpha_j) \\ &\geq \sum_{p=1}^i q_p H_j(z^p) \end{aligned}$$

Thus from Lemma 5.7 and Theorem 5.6 we may deduce that the corresponding departure processes satisfy $\mathbf{D}^0 \geq \sum_{i=1}^i q_p \mathbf{D}^p$.

For the case of round-robin routing, the routing functions as defined at the beginning of this example are given by $H_2(i) = \lceil i/2 \rceil$, and $H_3(i) = \lfloor i/2 \rfloor$. Note that the functions $H_2(i) = i/2 + 1/2$, and $H_3(i) = i/2$ also lead to the same constraint set \mathcal{T} . Further, this latter choice of functions gives us a stronger result as (5.12) is satisfied with

$$\beta_1 = \beta_2 = \frac{1}{2}, \quad -\epsilon_2 = \alpha_2 = \frac{1}{2}, \quad \text{and} \quad -\epsilon_3 = \alpha_3 = 0$$

Example 5.9 (Open Network with Splitting and Merging)
Consider the network shown in Fig. 4 that was used in the previous example, with the modification that there is an unlimited supply of jobs at station 1, i.e. $J_1 = \infty$. This corresponds to an open network where the first server models the exogenous arrival process. The initial configuration of jobs in this open network is given by $\mathbf{J} = (J_2, J_3, J_4)$ and the constraint set is defined by

$$\begin{aligned} \mathcal{T} = \left\{ \mathbf{x} \in \mathbb{Z}_+^4 : & i_2 \leq H_2(i_1) + J_2 \\ & i_3 \leq H_3(i_1) + J_3 \\ & i_4 \leq i_2 + i_3 + J_4 \right\} \end{aligned}$$

Now consider a synchronized family $\{\Delta(J^p) : p = 0, 1, \dots, i\}$ of such systems parameterized by their initial config-

urations. Let $q \in M^r$ be such that³

$$J^0 \geq \left\lceil \sum_{p=1}^r q_p J^p \right\rceil. \quad (5.14)$$

It is easy to see that whenever (x^1, \dots, x^r) are coeval, and $x^0 := \left\lceil \sum_{p=1}^r q_p x^p \right\rceil$, then $x_1^0 = x_1^1 = \dots = x_1^r$. Moreover, for $j = 2, 3$

$$\begin{aligned} x_j^0 &= \left\lceil \sum_{p=1}^r q_p x_j^p \right\rceil \\ &\leq \left\lceil \sum_{p=1}^r q_p (H_j(x_1^p) + J_j^p) \right\rceil \\ &= H_j(x_1^0) + \left\lceil \sum_{p=1}^r q_p J_j^p \right\rceil \\ &\leq H_j(x_1^0) + J_j^0. \end{aligned}$$

It is easy to see that x_1^0 also satisfies the required condition. Thus, Assertion (1) of Theorem 4.1 is satisfied, and hence

$$D^0(t) \geq \sum_{p=1}^r q_p D^p(t), \quad \forall t \geq 0. \quad (5.15)$$

Moreover, we also have

$$D_1^0(t) = D_1^1(t) = \dots = D_1^r(t), \quad \forall t \geq 0. \quad (5.16)$$

The total number of jobs, $N(t)$, in the system comprising of stations 2, 3, and 4, at time $t \geq 0$, is given by

$$N(t) = J_2 + J_3 + J_4 + D_1(t) - D_4(t). \quad (5.17)$$

Now if we assume that $J_2^p + J_3^p + J_4^p = J$ (const.) for $p = 0, 1, \dots, r$, then by (5.15), (5.16), and (5.17), we get

$$\begin{aligned} N^0(t) &= J_2^0 + J_3^0 + J_4^0 + D_1^0(t) - D_4^0(t) \\ &\leq J_2^0 + J_3^0 + J_4^0 + D_1^0(t) - \sum_{p=1}^r q_p D_4^p(t) \\ &= \sum_{p=1}^r q_p (J_2^p + J_3^p + J_4^p) + \sum_{p=1}^r q_p D_1^p(t) - \sum_{p=1}^r q_p D_4^p(t) \\ &= \sum_{p=1}^r q_p N^p(t). \end{aligned}$$

This establishes that the number of jobs in the system is convex in the initial configuration (J_2, J_3, J_4) under the constraint that $J_2 + J_3 + J_4 = J$ (const.).

The results obtained above are pathwise. We now consider a probabilistic routing mechanism that is symmetric, i.e.,

$$\{I_2(x); x \geq 1\} \stackrel{D}{=} \{I_3(x); x \geq 1\}. \quad (5.18)$$

Also assume that the servers at stations 2 and 3 are identical, and $J_4^1 = J_4^2$, $J_2^1 = J_3^2$, $J_3^1 = J_2^2$. Then because of the symmetry

$$\{N^1(t); t \geq 0\} \stackrel{D}{=} \{N^2(t); t \geq 0\}. \quad (5.19)$$

³Note that (5.14) is a strictly weaker condition than (5.13) which is what we would have required if we had not made use of coequality.

Thus, for the (symmetric) system with a symmetric routing mechanism and identical servers at stations 2 and 3, $E[N^J(t)]$, $t \geq 0$, is convex and symmetric in (J_2, J_3) with $J_2 + J_3 = \text{const.}$ and $J_4 = \text{const.}$. Thus $E[N^J(t)]$, $t \geq 0$, is increasing in $|J_2 - J_3|$, and therefore a more balanced load is better than a less balanced one (in the sense of majorization). So if at any time one has the option of "jockeying" jobs between queues 2 and 3, it is better to transfer jobs from the longer queue to the shorter one.

APPENDIX

PROOF OF PROPOSITION 5.2

1) α is rational, i.e., $\alpha_p = \frac{\nu_p}{\delta_p}$ where $\nu_p, \delta_p \in \mathbb{Z}_+$ are co-prime, $p = 1, \dots, r$. Let $\lambda = \text{l.c.m.}\{\delta_1, \dots, \delta_r\}$. Then, it is easy to see that $\text{g.c.d.}\{\frac{\lambda\nu_1}{\delta_1}, \dots, \frac{\lambda\nu_r}{\delta_r}\} = 1$. (If the above g.c.d. is $\gamma > 1$, then the l.c.m. will be $\frac{\lambda}{\gamma}$). Now consider the set

$$D := \left\{ \sum_{p=1}^r \frac{\lambda\nu_p}{\delta_p} x^p : x^p \in \mathbb{Z}_+, p = 1, \dots, r \right\}.$$

Clearly, D is closed under addition, and $\text{g.c.d.}(D) = 1$. Then, it follows that D contains all sufficiently large positive integers (see [22, Lemma 1-66]). Thus there exists $k, x^1, \dots, x^r \in \mathbb{Z}_+$, such that $\sum_{p=1}^r \frac{\lambda\nu_p}{\delta_p} x^p = k\lambda + 1$. This implies that $\sum_{p=1}^r \frac{\nu_p}{\delta_p} x^p = k + \frac{1}{\lambda}$. Thus

$$\left\lceil \sum_{p=1}^r \frac{\nu_p}{\delta_p} x^p \right\rceil - \sum_{p=1}^r \frac{\nu_p}{\delta_p} x^p = 1 - \frac{1}{\lambda}. \quad (A.1)$$

Moreover, for any $x^1, \dots, x^r \in \mathbb{Z}_+$, we have $\sum_{p=1}^r \frac{\lambda\nu_p}{\delta_p} x^p \in \mathbb{Z}_+$, and hence

$$\left\lceil \sum_{p=1}^r \frac{\nu_p}{\delta_p} x^p \right\rceil - \sum_{p=1}^r \frac{\nu_p}{\delta_p} x^p \leq 1 - \frac{1}{\lambda}. \quad (A.2)$$

That $s(\alpha) = 1 - \frac{1}{\lambda}$, follows by combining (A.1) and (A.2).

2) α is irrational. Assume, WLOG, that α_1 is irrational. It suffices to show that

$$\sup\{\lceil \alpha_1 x \rceil - \alpha_1 x : x \in \mathbb{Z}_+\} =: s = 1. \quad (A.3)$$

Since α_1 is irrational, (A.3) is equivalent to showing that

$$\inf\{\alpha_1 x - \lfloor \alpha_1 x \rfloor : x \in \mathbb{Z}_+, x \geq 1\} =: c = 0. \quad (A.4)$$

Let $C := \{\alpha_1 x - \lfloor \alpha_1 x \rfloor : x \in \mathbb{Z}_+, x \geq 1\}$. Now assume that $c > 0$. Let $l := \min\{k : kc \geq 1\}$. Note that $(l-1)c < 1 \leq lc < 1+c$. Let $\epsilon := 1+c-lc$. By the definition of c , we can choose a $d \in C$, such that $c \leq d < c + \frac{\epsilon}{l}$. It is easy to check that if $d \in C$, then $nd - \lfloor nd \rfloor \in C$ for any $n \in \mathbb{Z}_+, n \geq 1$. But $1 \leq lc \leq ld < lc + \epsilon = 1+c$, and thus $ld - \lfloor ld \rfloor = ld - 1 < c$. This gives us a contradiction, which establishes that $c = 0$. \square

REFERENCES

- [1] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages and Computation*. Reading, MA: Addison-Wesley, 1979.
- [2] F. Baccelli and Z. Liu, "Comparison properties of stochastic decision free Petri nets," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1905-1920, 1992.
- [3] P. Varaiya, "Finitely recursive processes," in *Discrete Event Systems: Models and Applications*. Sopron, Hungary, 1988, IIASA, Springer-Verlag, New York.
- [4] A. A. Lazar, "Optimal flow control of a class of queueing networks in equilibrium," *IEEE Trans. Automat. Contr.*, vol. AC-28, pp. 1001-1007, 1983.
- [5] J. G. Shanthikumar and D. D. Yao, "Second-order properties of the throughput of a closed queueing network," *Math. Oper. Res.*, vol. 13, no. 3, pp. 524-534, Aug. 1988.
- [6] J. G. Shanthikumar and D. D. Yao, "Second-order stochastic properties in queueing systems," *Proc. IEEE*, vol. 77, no. 1, pp. 162-170, Jan. 1989.
- [7] D. Mitra and I. Mitrani, "Analysis of a kanban discipline for cell coordination in production lines," *Management Sci.*, vol. 36, no. 12, pp. 1548-1566, Dec. 1990.
- [8] L. Meester and J. G. Shanthikumar, "Concavity of throughput and optimal buffer space allocation for tandem queueing systems with finite buffer storage space," *Adv. Appl. Prob.*, vol. 22, pp. 764-767, 1990.
- [9] V. Anantharam and Tsoucas, "The optimal buffer allocation problem," *Adv. Appl. Prob.*, pp. 761-764, 1990.
- [10] Y. C. Ho, M. A. Eyler, and T. T. Cien, "A gradient technique for general buffer storage design in a production line," *Int. J. Production Res.*, vol. 17, no. 6, pp. 557-580, 1979.
- [11] F. S. Hillier, K. C. So, and R. W. Boling, "Notes: Toward characterizing the optimal allocation of storage space in production line systems with variable processing times," *Management Sci.*, vol. 39, no. 1, pp. 126-133, 1993.
- [12] Y. Dallery and D. Towsley, "Symmetry property of the throughput in closed tandem queueing networks with finite buffers," *Oper. Res. Lett.*, vol. 10, pp. 541-547, 1991.
- [13] R. Rajan and R. Agrawal, "Cyclic networks with general blocking and starvation," *Queueing Systems: Theory and Applications*, Feb. 1994.
- [14] ———, "Second-order properties of families of discrete event systems," in *Proc. 32nd IEEE Conf. Decis. Contr.*, San Antonio, TX, Dec. 1993.
- [15] A. Ephremides, P. Varaiya, and J. Walrand, "A simple dynamic routing problem," *IEEE Trans. Automat. Contr.*, vol. AC-25, no. 4, pp. 690-693, Aug. 1980.
- [16] P. Glasserman and D. D. Yao, "Second-order properties of generalized semi-Markov processes," *Math. Oper. Res.*, vol. 17, pp. 444-469, 1992.
- [17] A. Ichikawa and K. Hiraishi, "Analysis and control of discrete event systems represented by Petri nets," in *Discrete Event Systems: Models and Applications*. Sopron, Hungary, 1988, IIASA, Springer-Verlag, New York.
- [18] F. Baccelli, G. Cohen, and B. Gaujal, "Recursive equations and new properties of timed Petri nets," *DEDS: Theory and Applications*, vol. 1, pp. 415-439, 1992.
- [19] A. Björner, L. Lovász, and P. Shor, "Chip-firing games on graphs," *Europ. J. Combinatorics*, vol. 12, pp. 283-291, 1991.
- [20] P. Glasserman and D. D. Yao, "Monotonicity in generalized semi-Markov processes," *Math. Oper. Res.*, vol. 17, pp. 1-21, 1992.
- [21] ———, "Structured buffer-allocation problems in production lines," preprint, December 1992.
- [22] J. G. Kemeny, J. L. Snell, and A. W. Knapp, *Denumerable Markov Chains*. Springer-Verlag, 1976, pp. 37-38.



Rajandran Rajan received the B.S. degree in mathematics from the University of Madras, the M.S. degree in statistics from the Indian Statistical Institute, New Delhi, and the M.S. degree in mathematics from the University of Wisconsin, Madison.

He is currently pursuing the Ph.D. degree in electrical engineering from the University of Wisconsin, Madison. His current research interests include the control of queueing systems and the analysis and control of timed discrete-event dynamical system as well as their continuous analogs.



Rajeev Agrawal (M'89) was born on December 1, 1963 in Lucknow, India. He received the B.Tech. degree in electrical engineering in 1985 from the Indian Institute of Technology, Kanpur and the M.S. and Ph.D. degrees in electrical engineering-systems from the University of Michigan, Ann Arbor, in 1987 and 1988, respectively.

Dr. Agrawal joined the University of Wisconsin-Madison in 1988, where he is currently an Associate Professor of Electrical and Computer Engineering. His current research interests include stochastic systems, stochastic adaptive control, resource allocation problems, communication networks, and queueing networks.

Technical Notes and Correspondence

Regional Pole Placement of Multivariable Systems Under Control Structure Constraints

S. Sathiya Keerthi and Makarand S. Phatak

Abstract—Many controller realizations are structurally constrained. Some typical examples are static output feedback, constant gain feedback for multiple operating points of a system, two-controller feedback, and decentralized feedback. A general class of problems of regional pole placement of multivariable systems with such control structure constraints is considered and a unified numerical method is given to solve them. First a problem in this class is converted to a problem of solving a system of equalities and inequalities. This system is then solved by using a modified homotopy method.

I. INTRODUCTION

Consider structurally constrained controllers, such as static output feedback, constant gain feedback for multiple operating points of a system, two-controller feedback, and decentralized feedback. With such constrained controllers it is in general not possible to place the eigenvalues arbitrarily in the complex plane. For satisfactory dynamical behavior of a system, it usually suffices to place the eigenvalues in some desired stability region, S in the complex plane, i.e., to S -stabilize the system. With structurally constrained controllers S -stabilization is quite possible. In this paper we consider a general class of problems of S -stabilization (or regional pole placement) of multivariable systems with control structure constraints and give a numerical method to solve them. This work is a nontrivial extension of the authors' earlier work [6] to the multivariable case. While doing the extension the possible overlap between [6] and this note is kept at a minimum by omitting the most repetitive details.

The problems mentioned here have also been considered in many earlier papers from which we choose references somewhat arbitrarily and point to [3] for the case of static output feedback, [8] for the case of constant gain feedback for multiple operating points of a system, [9] for the case of two-controller feedback, and [5] for the case of decentralized feedback. Our method is fundamentally different from the above methods and treats different problems under a single framework.

In our method we formulate (Section II) each of the regional eigenvalue placement problems as a problem of solving a system of equalities and inequalities, which we solve (Section III) by employing a modified homotopy framework. An example is presented (Section IV) to show the efficacy of our method.

II. PROBLEM FORMULATION

Consider the multivariable dynamical system

$$\dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t) \quad (2.1)$$

Manuscript received June 16, 1992; revised January 12, 1993, June 8, 1993, October 20, 1993, and March 2, 1994.

S. S. Keerthi is with the Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India.

M. S. Phatak was with the Department of Computer Science and Automation, Indian Institute of Science and is now with Aerospace Systems Private Limited, Bangalore, India.

IEEE Log Number 9405662.

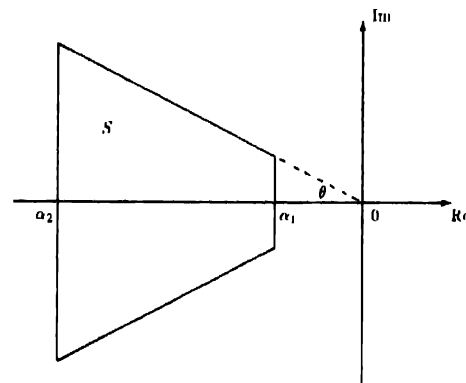


Fig. 1. An example of a popular stability region S in the complex plane.

where $x \in R^n$, $u \in R^m$, $y \in R^p$, and A , B , C are constant matrices of appropriate dimensions. We will assume the following throughout: system (2.1) is controllable and observable, $\text{rank } B = m$ and $\text{rank } C = p$. The feedback we consider is of the form

$$u(t) = -Kx(t). \quad (2.2)$$

Our objective is to S -stabilize, i.e., to place the eigenvalues of $(A - BK)$ in some desired stability region, S in the complex plane under given control structure constraints. S comes from specifications of the desirable dynamical behavior of the closed-loop system. In particular we focus our attention on the popular stability region, S shown in Fig. 1. Control structure constraints are imposed by the way feedback is realized. For example, in the case of static output feedback the constraint is that rows of K must belong to the range space of rows of C so that state feedback (2.2) can be realized as static output feedback. To solve this constrained problem we formulate a system of inequalities and equalities

$$g(v) \leq 0 \quad (2.3)$$

$$h(v) = 0 \quad (2.4)$$

where v is some set of intermediate variables. It is worth mentioning here that a neat transformation of our control problem to an instance of (2.3)–(2.4) is by no means trivial; so our way of doing it is interesting in its own right. Once a v satisfying (2.3)–(2.4) is found the gain matrix required for stabilization can easily be found. In Section II-A we present details of how v is chosen and how to formulate the inequalities. The equalities will be derived in Section II-B.

A. Choice of Variables and Formation of Inequalities

Given a controllable pair (A, B) the feedback matrix K which gives a specified characteristic polynomial for $(A - BK)$ is not unique. K , however, is determined uniquely by a specified matrix polynomial $P(s)$ [1]. $P(s)$ is defined by

$$P(s) = \bar{P}(s) + \Gamma \bar{\Gamma}^*(s) \quad (2.5)$$

where Γ is a constant (i.e., independent of s) $m \times n$ matrix

$$\bar{P}(s) = \text{diag}[s^{n_1}, \dots, s^{n_m}].$$

$$V(s) = \text{block diag}[(1 \ s \ -a^{n_1-1})^T, \dots, (1 \ s \ -a^{n_l-1})^T]$$

and n_1, \dots, n_l are the controllability indexes of the pair (A, B) he K which satisfies

$$\det(sI - A + BK) = \det(P(s)) \quad (2.6)$$

is given by

$$K = I^{-1}(\bar{K} + \Gamma T) \quad (2.7)$$

where the matrices I (nonsingular and upper triangular), \bar{K} and T (nonsingular) depend on (A, B) see [1] for details of the dependency. Since the relationship between K and I is affine as given by (2.7) Γ is an attractive choice for the set of variables, γ . However the problem of finding conditions on the elements of Γ so that the set of eigenvalues, $\Lambda = \lambda(A - BK) \subset S$ is very hard. On the other hand suppose we denote

$$\det(P(s)) = \det(P(s) + \Gamma V(s)) = \left[\prod_{i=1}^l (s + a_i s + b_i) \right] (s - \lambda) \quad (2.8)$$

where $l = \lfloor n/2 \rfloor$ the integer part of $n/2$ and the last term $(s - \lambda)$ occurs when n is odd. Then the following proposition shows that the constraint $\lambda(A - BK) \subset S$ can be easily reformulated in terms of the elements of γ where

$$- [a_1 b_1 \dots a_l b_l \lambda]^T \quad (2.9)$$

Proposition 2.1 Let S be as shown in Fig. 1 then $\lambda(A - BK) \subset S$ is equivalent to

$$\alpha_+ \leq -a/2 \leq \alpha_- \quad b \leq (1 + \tan \theta)(a/2)^+$$

$$\alpha_+ + \alpha_+ a + b \geq 0 \quad \alpha_- + \alpha_- a + b \geq 0 \quad \forall i \in \{1, \dots, l\}$$

$$\alpha_+ \leq \lambda \leq \alpha_- \quad (2.10)$$

By using the results on representation of general stability regions in terms of inequalities [1 p. 298–299] Proposition 2.1 can be easily proved. We omit the details for the sake of brevity.

The use of quadratic factors in (2.8) avoids the need to work with complex numbers. We define the set of variables γ by

$$\gamma = \{\Gamma\} \quad (2.11)$$

It is clear that the elements of γ are not independent. The reason for choosing such an overdetermined set, γ , is given later in Remark 2.2. In Section II-B we develop equality constraints which must be satisfied by the elements of Γ and γ to ensure (2.6) and (2.8). The system of inequalities (2.10) define $q(\gamma) \leq 0$ (truly $q(\gamma) \leq 0$) in (2.3).

B. Equalities

We first formulate the equalities which are directly associated with the definition of the set of variables, γ in (2.11).

Assumption 2.1 Let $p_i(s) = s^2 + a_i s + b_i$, $i = 1, \dots, l$, and $p_{l+1}(s) = s - \lambda$. For each $i, j \in \{1, \dots, l+1\}$, $i \neq j$, $p_i(s)$ and $p_j(s)$ are relatively prime.

Proposition 2.2 Suppose that Assumption 2.1 holds. It is equivalent to the system of n equalities given by

$$\epsilon_1 [\text{block det}(\bar{P}(C) + \Gamma V(C))] = 0$$

$$\det(\bar{P}(\lambda) + \Gamma V(\lambda)) = 0$$

where

$$\epsilon_1 = [1 \ 0] \quad C = \begin{bmatrix} 0 & 1 \\ -1 & -a \end{bmatrix} \quad \text{and}$$

$$\text{block det}(P(C)) = \det(P(s))$$

Proof. We can eliminate s in (2.8) by substituting it by C , $i = 1, 2, \dots, l$ and λ so that the right-hand side of (2.8) is zero by Cayley-Hamilton's theorem. Then we get an intermediate system of equations which is the same as (2.12) except that the term ϵ_1 is absent. This intermediate system of equations contains a total of $4l+1 = (n+2l)$ equations. Out of these $2l$ are redundant and when they are eliminated we get (2.12). Thus (2.8) implies (2.12).

We now show using Assumption 2.1 that (2.12) implies (2.8). Let $p_i(s)$ be as in Assumption 2.1. If we show that for each $i = 1, \dots, l+1$

$$p_i(s) \text{ divides } \det(\bar{P}(s) + \Gamma V(s)) \quad (2.13)$$

then we are done because of Assumption 2.1 and the fact that the degree of $\det(\bar{P}(s) + \Gamma V(s))$ is equal to n . By the second level equation in (2.12) it directly follows that (2.13) holds for $i = l+1$. Now take $i \in \{1, \dots, l\}$. To show (2.13) we need to consider two cases: (1) $\lambda_1 = \lambda$, (2) $\lambda_1 \neq \lambda$ where λ_1 and λ are the roots of $s^2 + a_i s + b_i = 0$.

In case 1, $s^2 + a_i s + b_i = 0$ has a root λ_1 of multiplicity two. We have

$$C = \begin{bmatrix} 1 & 0 \\ \lambda_1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \lambda_1 & 1 \end{bmatrix}^{-1}$$

and

$$\begin{aligned} & \text{block det}(\bar{P}(C) + \Gamma V(C)) \\ &= \begin{bmatrix} 1 & 0 \\ \lambda_1 & 1 \end{bmatrix} \\ & \quad \left[\begin{array}{cc} \det(\bar{P}(\lambda_1) + \Gamma V(\lambda_1)) & \frac{1}{\lambda_1} \det(\bar{P}(\lambda_1) + \Gamma V(\lambda_1)) \\ 0 & \det(\bar{P}(\lambda_1) + \Gamma V(\lambda_1)) \end{array} \right] \\ & \quad \begin{bmatrix} 1 & 0 \\ \lambda_1 & 1 \end{bmatrix}^{-1} \end{aligned} \quad (2.14)$$

From (2.12) and (2.14) it follows that

$$\det(\bar{P}(\lambda_1) + \Gamma V(\lambda_1)) = 0 \quad \frac{d}{d\lambda_1} \det(\bar{P}(\lambda_1) + \Gamma V(\lambda_1)) = 0$$

and hence (2.13) holds. A similar (actually easier) proof holds for case 2. This completes the proof of the proposition. ■

Remark 2.1 Equations in (2.12) are independent. The proof of this is implicit in Lemma 3.2 which is stated and proved in Section III.

Remark 2.2 In the single-input case K is determined uniquely by γ and can be expressed directly in terms of the elements of γ as [using factorization of the scalar polynomial $P(s)$ in (2.8)] $K = q \cdot [\prod_{i=1}^l (s^2 + a_i s + b_i I)](sI - A)^{-1}$, where q is as defined in [1]. So (2.12) is not required in the single-input case [6]. In the multi-input case K is determined uniquely by Γ and cannot be expressed directly in terms of the elements of γ and some other variables required for its unique determination. This is because of the lack of a factorized parameterization of the matrix polynomial $P(s)$. So we define the overdetermined set γ in (2.11) and the constraining equations (2.12) for the multi-input case.

The block determinant in (2.12) and its partial derivatives with respect to the elements of v , which are required in the numerical solution described in Section III, can be obtained by using a transformation of a polynomial matrix to Hermite form [2]. This transformation uses elementary operations to reduce a square polynomial matrix to triangular form. First $\det(P(s))$ is obtained from the Hermite form of $P(s)$ as $\det(P(s)) = s^n + \beta_1 s^{n-1} + \dots + \beta_n$. Then block $\det P(C)$ is obtained by evaluating the above polynomial at C , i.e., (subscript i in C , a , and b is dropped in the following expressions for notational simplicity) block $\det P(C) = \det(P(s))|_{s=C} = C^n + \beta_1 C^{n-1} + \dots + \beta_n I_2$. Partial derivatives of block determinant with respect to elements of Γ are obtained as follows. From $P(s)$ equal to the expression found at the bottom of the page where γ_i^{jk} 's are the elements of Γ ; the partial derivative of block $\det P(C)$ with respect to, say, γ_1^{11} , is nothing but $C^{n-1}[\text{adj } P(s)]_{11}|_{s=C}$, where $[\text{adj } P(s)]_{11}$ is the (1, 1) element of the adjoint of $P(s)$. The partial derivatives with respect to other γ_i^{jk} 's are similarly obtained. Partial derivatives of block determinant with respect to elements of C are obtained from the partial derivatives of power of C which are given below

$$\begin{aligned}\frac{\partial}{\partial a} C^n &= \sum_{j=1}^n C^{(j-1)} \frac{\partial}{\partial a} (C) C^{n-j}, \\ \frac{\partial}{\partial b} C^n &= \sum_{j=1}^n C^{(j-1)} \frac{\partial}{\partial b} (C) C^{n-j}, \quad 1 \leq r \leq n\end{aligned}$$

where

$$\frac{\partial}{\partial a} (C) = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}, \quad \frac{\partial}{\partial b} (C) = \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}.$$

Remark 2.3: Assumption 2.1 can be enforced by simple inequalities, using 2×2 McDuffe resultants [2]. These inequalities along with (2.10) actually define $g(v) \leq 0$ in (2.3). In the actual implementation, however, we have preferred to omit these extra inequalities arising from enforcing Assumption 2.1. We have also observed, on all the examples tried, that this omission does not lead to any difficulties in obtaining the final solution. We will discuss more about this in Section III (Remark 3.1).

Remark 2.4: An important observation is worth mentioning here. Our algorithm for solving the constrained stabilization problems, which will be discussed later, starts from an initial guess point and varies a , and b , continuously, while satisfying (2.12), so as to reach a solution. Since (2.12) allows a pair of repeated roots, a continuous change between a pair of real roots and a pair of complex conjugate roots is possible.

Now we consider the equalities which arise from structural constraints on state feedback. By appropriately appending (2.12) to these equalities we get (2.4). We demonstrate the setting up of the equalities only for the control structure of decentralized state feedback. Details for the other control structures mentioned earlier are straightforward and can be found in [6]. The K given by (2.7) is function of Γ , a part of v , and so we denote K by $K(v)$. The control structure constraints on K are treated as constraints on v .

Decentralized State Feedback

Let r be the number of subsystems of a large scale system for which a decentralized feedback is to be realized. The state feedback gain matrix, K typically takes the block diagonal form: $K = \text{block diag}[K_1, \dots, K_r]$. The control structure constraints corresponding to this form can be expressed as

$$E_i^L K(v) E_i^R = 0, \quad i = 1, 2, \dots, r \quad (2.15)$$

where the matrices E_i^L, E_i^R are constant matrices of appropriate dimensions which select all blocks in the i th block row except the i th one. Sometimes there exists an overlap between the state variables of the subsystems. In such a case K cannot be expressed in block diagonal form. The structural constraints can still be represented as in (2.15), however, with an appropriate choice of E_i^L and E_i^R , $i = 1, \dots, r$. Together (2.15) along with (2.12) define (2.4).

Remark 2.5: The equations in (2.12) involve characteristic polynomials and determinants, quantities which are known to have bad sensitivity properties. In other words, in the presence of finite precision arithmetic, the difference between the set of eigenvalues of $(A - BK)$ and the set of roots of (2.8) can be large even when the errors in the satisfaction of (2.12) are small. Therefore, when a K is determined using the numerical method proposed in this paper, it is necessary to evaluate the eigenvalues of $(A - BK)$ and check whether they belong to S . If this check fails then it indicates that our method has suffered from numerical instability. If computations are done with a high precision this would not happen. See for instance, the example of Section IV.

III. A MODIFIED HOMOTOPY METHOD

We solve the system of equalities and inequalities obtained in the previous section by using a modified homotopy method. Homotopy methods are popularly used to find zeros of a square system of non linear equations. Our main reason for preferring homotopy methods over the usual iterative methods is that it has been observed in practice that the domain of attraction of a solution point for iterative methods is usually much smaller than that for homotopy methods. A good survey of homotopy (or, continuation) methods together with guiding references on theory, numerical methods and applications is given in [7]. We devise a modification of a homotopy method to solve the system of equalities (2.4) (which is typically underdetermined) along with the inequalities (2.3).

Let $H(v, t)$ be a function with the following properties: a) when $t = 0$, the system of equations $H(v, 0) = 0, g(v) \leq 0$ is trivial in the sense that it is easy to find a \bar{v} satisfying them, and b) when $t = 1$, $H(v, 1) = h(v)$. Thus when $t = 0$ we know that \bar{v} is a solution to $H(v, 0) = 0, g(v) \leq 0$. In a homotopy method one starts from this known solution $(\bar{v}, 0)$, and moves on the solution hypersurface defined by $\{(v, t): H(v, t) = 0, g(v) \leq 0, 0 \leq t \leq 1\}$ in an attempt to reach a solution v^* , if it exists. Our choice of the H function is

$$H(v, t) = \begin{bmatrix} h^1(\Gamma) + (t-1)h^1(\bar{\Gamma}) \\ h^2(\Gamma, \gamma) \end{bmatrix} \quad (3.1)$$

where: the h^1 function represents the given control structure constraint, see for example, (2.15) for the case of decentralized state

$$\begin{bmatrix} s^{n_1} + \gamma_1^{11} s^{n_1-1} + \dots + \gamma_{n_1}^{11} & \gamma_1^{12} s^{n_1-1} + \dots + \gamma_{n_1}^{12} & \dots & \gamma_1^{1m} s^{n_1-1} + \dots + \gamma_{n_1}^{1m} \\ \gamma_1^{21} s^{n_1-1} + \dots + \gamma_{n_1}^{21} & s^{n_2} + \gamma_1^{22} s^{n_2-1} + \dots + \gamma_{n_2}^{22} & \dots & \gamma_1^{2m} s^{n_1-1} + \dots + \gamma_{n_2}^{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_1^{m1} s^{n_1-1} + \dots + \gamma_{n_1}^{m1} & \gamma_1^{m2} s^{n_2-1} + \dots + \gamma_{n_2}^{m2} & \dots & s^{n_m} + \gamma_1^{mm} s^{n_m-1} + \dots + \gamma_{n_m}^{mm} \end{bmatrix}$$

feedback; the h^2 function is the left-hand side of (2.12); and $\bar{\Gamma}$ is chosen as explained below. Recall that $v = \{\Gamma, \gamma\}$. When $t = 0$, we obtain a solution \bar{v} to $H(v, 0) = 0$ as follows. Choose a $\bar{\gamma}$ so that $g(\bar{\gamma}) < 0$. Note that it is easy to get such a $\bar{\gamma}$ because of the decoupled nature of the inequalities; see (2.10). Once a $\bar{\gamma}$ is chosen, choose a $\bar{\Gamma}$ so that $h^2(\bar{\Gamma}, \bar{\gamma}) = 0$. Again, it is easy to choose such a $\bar{\Gamma}$; for example one possibility is to choose $\bar{\Gamma}$ which corresponds to a diagonal matrix $P(s)$. Then, \bar{v} is nothing but $\{\bar{\Gamma}, \bar{\gamma}\}$.

To move on the solution hypersurface from $t = 0$ to $t = 1$ we devise a numerical method by using an optimization set-up. Here our description of the optimization set-up will be very brief; see [6] for its detailed discussion and the algorithm based on it.

Since $t = 1$ has to be reached for the homotopy method to be successful, and $t = 1$ has to be reached before crossing over to $t > 1$, $(1 - t)$ is a good choice for the objective function to be minimized. So we define the following optimization problem

$$\min(1 - t) \text{ subject to } H(v, t) = 0, \quad g(v) \leq 0. \quad (3.2)$$

It is important to note that, with $(1 - t)$ as the cost function, the aim is not to solve the optimization problem (3.2). It is formulated only to drive the motion from $t = 0$ to $t = 1$. It is easy to maintain feasibility with respect to the inequalities $g(v) \leq 0$ simply by using a barrier method [4] and replacing (3.2) by

$$\min(1 - t) + c_k B(v) \text{ subject to } H(v, t) = 0 \quad (3.3)$$

where $B(v)$ is a barrier function which is smooth and which tends to ∞ , i.e., establishes a barrier, as v approaches the boundary of $\{v: g(v) \leq 0\}$ and c_k is positive. Problem (3.3) is repeatedly solved for a sequence of c_k values (which monotonically decrease to zero). This solution process is terminated as soon as $t = 1$ is reached and there is no need to go for an actual minimization of (3.2). Therefore, unlike usual barrier methods, our method requires that the solution of (3.3) is repeated only for a few c_k values. In our numerical implementation we have used a logarithmic barrier function [4]. See [6] for the rationale used to choose the decreasing sequence $\{c_k\}$. In all the test examples it was observed that the initial choice $c_1 = 0.1$ worked well in that it was sufficient to solve (3.3) only once, i.e., $t = 1$ was reached while solving (3.3) with c_1 .

Suppose that the S -stabilization problem does not have a solution. In that case the method will reach a stage where the solution trajectory comes close to making one or more of the inequalities active and, as a result t gets stuck at a value less than one. In that case the region S can be expanded by appropriately changing the active inequalities and the procedure can be restarted.

Problem (3.3) is only an equality constrained problem. We solve (3.3) by a continuous realization of the gradient projection method [4, 6]. Let

$$Z = \{(v, t): H(v, t) = 0\} \quad (3.4)$$

and $\nabla f_p(v, t)$ be the projection of the gradient of $f = (1 - t) + c_k B(v)$ onto the tangent space of Z at (v, t) . In the gradient projection method one moves on Z along the trajectory of the vector field defined by $-\nabla f_p$. Careful tracking of the solution manifold Z using the vector field $-\nabla f_p$ can be done by using a simple modification of any state-of-the-art ODE solving package [10]. Such a tracking allows continuous check on any deviation from the solution manifold caused by the propagation and accumulation of numerical errors and applies a correction to get back to the solution manifold whenever necessary.

Let C denote the curve on Z starting from $(\bar{v}, 0)$ and defined using the above process. The tracking of C requires Z to be smooth. The smoothness of H does not automatically guarantee smoothness of Z . The following result describes the smoothness of Z .

Theorem 3.1: Let Z be as defined in (3.4) and assume that $\{(v, t): \text{Assumption 2.1 holds}\} \subset R^N$. Suppose that the Jacobian of h^1 with respect to Γ , has full row rank for every $(v, t) \in Z \cap V$. Then $Z \cap V$ is a differentiable manifold.

Remark 3.1: Typically C cuts ∂V , the boundary of V , transversally and so there is really no need to worry about any difficulties imposed by any crossing of ∂V (if at all it occurs) during the tracking of C . In doing all the numerical examples we have simply ignored Assumption 2.1 and yet did not face any difficulties in reaching a solution $(v^*, 1)$.

Remark 3.2: The assumption that h^1 has full row rank usually comes directly from the way the control structure constraints are formulated.

To prove Theorem 3.1 we use the following lemma.

Lemma 3.2: Suppose v is such that Assumption 2.1 holds and $h^2(v) = h^2(\Gamma, \gamma) = 0$, where h^2 is as in (3.1). Then $h^2_\gamma(\Gamma, \gamma)$, the Jacobian of h^2 with respect to γ , is nonsingular.

Proof: Let $\delta(s) = \det(\bar{P}(s) + \Gamma V(s))$. With straightforward algebra it is easy to see that $h^2_\gamma(\Gamma, \gamma)$ has block diagonal structure with i th ($i \in \{1, \dots, l\}$) block given by

$$\begin{bmatrix} \frac{\partial \delta(C_i)}{\partial a_i} \\ \frac{\partial \delta(C_i)}{\partial b_i} \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \prod_{j=1, j \neq i}^{l+1} G_j(C_i),$$

$$G_j(C_i) = C_i^2 + a_j C_i + b_j I_2, \quad j = 1, \dots, l$$

$$G_{l+1}(C_i) = C_i - \lambda I_2.$$

The last $((l+1)$ th) block is $\partial \delta(\lambda)/\partial \lambda = \prod_{i=1}^l (\lambda^2 + a_i \lambda + b_i)$. If Assumption 2.1 holds then for each i, j , $G_j(C_i)$ (which is the McDuffe resultant associated with $p_i(s)$ and $p_j(s)$) is nonsingular [2] and $\lambda^2 + a_i \lambda + b_i$ is nonzero. Hence Lemma 3.2 follows. ■

Proof of Theorem 3.1: To prove Theorem 3.1 it is sufficient to show that the Jacobian J of H with respect (Γ, γ, t) has full row rank for every $(v, t) = (\Gamma, \gamma, t) \in Z \cap V$. We have

$$J = \begin{bmatrix} h^1_\Gamma & 0 & h^1(\bar{\Gamma}) \\ h^2_\Gamma(\Gamma, \gamma) & h^2_\gamma(\Gamma, \gamma) & h^2(\bar{\Gamma}, \bar{\gamma}) \end{bmatrix}. \quad (3.5)$$

Then from (3.5) and Lemma 3.2, Theorem 3.1 follows directly. ■

It is very important that C connects the points $(\bar{v}, 0)$ and $(v^*, 1)$ (if v^* exists). This connectivity condition depends on the choice of H . For any of the S -stabilization problems which are considered in this paper choosing an H such that the connectivity condition is satisfied is a very hard problem. We believe that it can only be solved when complete system theoretic solutions to these problems become available. Our choice of H given in (3.1) is not guaranteed to satisfy the connectivity condition. Tests on several problems have shown, however, that the homotopy method based on the H as in (3.1) and the optimization formulation (3.2) has good empirical success. On all the problems tried we were always successful in reaching $t = 1$ without encountering any local minima of (3.2) or any unboundedness on the elements of Γ .

IV. EXAMPLE

We carried out numerical tests on several constrained S -stabilization problems using the VAX-88 computer. Here we present only one example to demonstrate the effectiveness of our approach.

Power System: We consider the example of seven-state, two-input model of load frequency control of a two-area power system. The numerical data for the A, B matrices are taken from [5]. Here we also take into consideration integral control action which is employed to reject step disturbances due to changes in load. Thus we append the state vector by two more variables which represent the integrals of the area control errors defined as $\Delta p_{tie} + \Delta f_1$ and $-\Delta p_{tie} + \Delta f_2$.

TABLE I
SOLUTION FOR POWER SYSTEM

Initial values ($t = 0$)	
Λ	$\{-2 \pm j1, -3 \pm j1, -4 \pm j1, -5 \pm j1, -2\}$
$P(s)$	$p_{11}(s) = s^5 + 12s^4 + 58s^3 + 148s^2 + 190s + 100$ $p_{12}(s) = 0$ $p_{21}(s) = 0$ $p_{22}(s) = s^4 + 18s^3 + 123s^2 + 378s + 442$
$E_1^T K E_1^R$	$E_1^T K E_1^R = \begin{bmatrix} 0.984 & 1.127 & 0.480 & 0.517 \\ 0.234 & 0.078 & 0.000 & 0.884 \end{bmatrix}$ $E_2^T K E_2^R = \begin{bmatrix} 0.984 & 1.127 & 0.480 & 0.517 \\ 0.234 & 0.078 & 0.000 & 0.884 \end{bmatrix}$

Number of integration steps = 1954, cpu time = 736.15 sec

Feasible solution values ($t = 1$)	
Λ	$\{-1.47087 \pm j0.691058, -2.72493 \pm j0.992018,$ $-4.42945 \pm j1.49412, -6.27487 \pm j3.05654,$ $-1.99674\}$
$P(s)$	$p_{11}(s) = s^5 + 13.9234s^4 + 82.3513s^3 + 271.316s^2 + 338.132s + 234.420$ $p_{12}(s) = -3.95005s^3 - 40.9614s^2 - 101.740s - 108.530$ $p_{21}(s) = -3.27041s^3 - 58.2824s^2 - 10.7900s - 146.782$ $p_{22}(s) = s^4 + 17.8735s^3 + 116.771s^2 + 289.408s + 269.346$
$E_1^T K E_1^R$	$E_1^T K E_1^R = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ $E_2^T K E_2^R = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$
K	$(k_1, k_{1s}) = \begin{bmatrix} 0.238848 & 0.959565 & 1.84323 & 0.643260 & -0.580120 \end{bmatrix}$ $(k_{2s}, k_2) = \begin{bmatrix} 2.70128 & 0.637658 & 1.621895 & 2.159241 & 1.07738 \end{bmatrix}$

see [5] for the physical meanings of $-\Delta p_{i1}$ and Δf_i , $i = 1, 2$ and the other state variables. The feedback is decentralized with some overlap. The 2×9 gain matrix K has the following structure

$$K = \begin{bmatrix} \times & \times & \times & \times & \times & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \times & \times & \times & \times & \times \end{bmatrix}$$

$$= \begin{bmatrix} k_1 & k_{1s} & 0 \\ 0 & k_{2s} & k_2 \end{bmatrix}$$

where the \times 's represent the free gain elements. Results obtained by our method on this example are tabulated in Table I. We used the stability region of Fig. 1 with $\alpha_1 = -1$, $\alpha_2 = -10$ and $\theta = 45^\circ$. For doing homotopy curve tracking we used relative tolerance of 10^{-4} . In view of Remark 2.5 we compared the eigenvalues of $(A - BK)$ and the roots of (2.8). They matched up to three decimal digits.

V. CONCLUSION

In this note we have given a numerical method for solving a general class of multivariable regional pole placement problems with control constraints, which has worked well on many examples. Given that for general control structure constraints the problem is very hard, such a method should prove very useful.

REFERENCES

- [1] J. Ackermann, *Sampled-Data Control Systems, Analysis and Synthesis, Robust System Design*. Berlin: Springer-Verlag, 1985.
- [2] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1980.
- [3] L. H. Keel and S. P. Bhattacharyya, "State-space design of low-order stabilizers," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 182-186, 1990.
- [4] D. G. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1984.
- [5] A. K. Mahalanabis and R. Singh, "On decentralized feedback stabilization of large scale interconnected systems," *Int. J. Contr.*, vol. 32, pp. 115-126, 1980.
- [6] M. S. Phatak and S. S. Keerthi, "A homotopy approach for stabilizing single-input systems with control structure constraints," *Automatica*, vol. 28, pp. 981-987, 1992.

- [7] S. L. Richter and R. A. DeCarlo, "Continuation methods: theory and applications," *IEEE Trans. Automat. Contr.*, vol. AC-28, pp. 660-665, 1983.
- [8] W. E. Schmitendorf and C. Wilmers, "Simultaneous stabilization via low order controllers," in *Control and Dynamic Systems*, C. T. Leondes, Ed. New York: Academic, 1990, p. 165-184.
- [9] D. D. Siljak, "Dynamic reliability using multiple control systems," *Int. J. Contr.*, vol. 31, pp. 303-329, 1980.
- [10] H. A. Watts, "RDEAM—An Adams ODE code with root solving capability," SAND 85-1595, Sandia Laboratories, Albuquerque, NM, 1985.

Continuous Robust Control Design for Nonlinear Uncertain Systems Without a Priori Knowledge of Control Direction

J. Kaloust and Z. Qu

Abstract—In this paper, a robust control scheme is proposed for a class of nonlinear systems that have not only additive nonlinear uncertainties but also unknown multiplicative signs. These signs are called control directions since they represent effectively the direction of motion under any given control. Except for the unknown control directions, the class of systems satisfy the generalized matching conditions. Nonlinear robust control is designed to identify on-line unknown control directions and to guarantee global stability of uniform ultimate boundedness without the knowledge of nonlinear dynamics except their size bounding functions. It is also shown that the proposed robust control can be made continuous through utilizing the so-called shifting laws that change smoothly and accordingly the signs of robust controls and that, no matter what time constants and gains of the shifting laws are, global stability is always ensured. The analysis and design is done using Lyapunov's direct method.

1. INTRODUCTION

Robust control of nonlinear systems in the presence of nonlinear uncertainties has been studied extensively. The important classes of stabilizable uncertain systems and their robust control laws can be found in [1], [3], [6], [9]-[11]. The uncertainties in those systems can be general nonlinear functions and input-related and/or input-unrelated. The input-related uncertainties studied so far in the previous results, however, are not only sign-invariant but also have known signs. These signs, called control directions, represent motion directions of the system under any control, and knowledge of these signs makes robust control design much easier. The objective of this paper is to develop a robust control design procedure based on Lyapunov's direct method that achieve global stability but does not require a priori knowledge of control directions. For practical implementation, designs of both continuous and discontinuous robust controls are considered and compared. It is shown that continuity of robust control can be achieved by designing the so-called shifting laws that change the signs of the robust control in a continuous fashion. The shifting laws are based on the results of on-line identification of unknown control directions.

Manuscript received October 14, 1993; revised January 26, 1994 and May 2, 1994. This work was supported in part by National Science Foundation Grant MSS-9110034.

The authors are with the Department of Electrical Engineering, University of Central Florida, Orlando, FL 32816 USA.

IEEE Log Number 9407023.

It is worth mentioning that the problem of unknown control directions has been studied for the last 10 years in the area of adaptive control. Most results in that area are for linear time invariant systems with unknown parameter and unknown high-frequency gain (that is, control direction). The first result was proposed by Nussbaum [8] in which the adaptive control uses the so-called Nussbaum gain, a transcendental function whose sign changes an infinite number of times as its argument tends to infinity. Using the same principle in [8], Mudgett and Morse [7] developed adaptive control for general linear systems without knowledge of high-frequency gain. An alternative method called correction vector approach was proposed in [5]. Recently, efforts have been made to extend these results to nonlinear systems that have unknown control directions. In [2], [4] two adaptive control schemes have been proposed for simple first-order nonlinear systems.

Compared with these existing results (especially the two on nonlinear systems), the results proposed in this paper have several advantages. First, the class of systems studied here is second-order, and the basic idea and approach used can be extended to second-order vector and higher-order systems. Second, systems considered in this paper are general nonlinear systems with general nonlinear time-varying uncertainties, while those in [2], [4] are time-invariant, have only unknown constants, and satisfy the Lipschitz condition. Third, the robust control laws provided here can be selected to be continuous in contrast with discontinuous adaptive control schemes formulated in [2], [4]. Fourth, the proposed robust control scheme guarantees global stability of uniform ultimate boundedness in the presence of general nonlinear uncertainties. Finally, instead of stabilization problem formulated in [2], [4], output tracking problem is solved in this paper.

The outline of this paper is as follows. In Section II formulation of our robust control problem of nonlinear uncertain system is presented. Robust control and shifting laws are developed in Section III, followed by a simulation example. Finally, the conclusion is drawn in Section IV.

II. PROBLEM FORMULATION

We shall consider a second-order nonlinear uncertain system to be described by

$$\dot{x}_1 = f_1(x_1, t) + \Delta f_1(x_1, \eta_1, t) + g_1(x_1, a_1 x_2, \eta_1, t) \quad (1)$$

$$\begin{aligned} \dot{x}_2 = & f_2(x_1, x_2, t) + \Delta f_2(x_1, x_2, \eta_2, t) \\ & + g_2(x_1, x_2, a_2 u, \eta_2, t) \end{aligned} \quad (2)$$

where $x = [x_1 \ x_2]^T \in \mathbb{R}^2$ is the state, $u(\cdot) \in \mathbb{R}$ is the control input, the variables η_1 and η_2 represent bounded, uncertain time-varying parameters, and a_1 and a_2 are scaled (in the sense that $|a_1| = |a_2| = 1$), constant parameters which characterize directions of controlled motion. The control directions, $\text{sign}[a_1]$ and $\text{sign}[a_2]$, are unknown. The functions $f_1(x_1, t)$ and $f_2(x_1, x_2, t)$ are known, functions $\Delta f_1(x_1, \eta_1, t)$ and $\Delta f_2(x_1, x_2, \eta_2, t)$ represent uncertainties, and the functions $g_1(x_1, a_1 x_2, \eta_1, t)$ and $g_2(x_1, x_2, a_2 u, \eta_2, t)$ may contain uncertainty in addition to a_1 and a_2 as well.

The class of uncertain system is given by (1) and (2) is a special case of the class of nonlinear systems in [10]. These systems satisfy the structure specified by the so-called generalized matching conditions. The difference between [10] and this paper is that the above systems has unknown control directions. For simplicity of presentation, subsystems (1) and (2) are assumed to be scalar. Extension can be made to higher-order and vector systems in which the control directions a_i are diagonal and constant matrices of proper dimensions.

We introduce the following assumptions for system (1) and (2)

- A.1) Under a continuous control $u(x, t)$, system (1) has a classical solution.
- A.2) The uncertainties Δf_1 and Δf_2 are bounded, continuous, and locally uniformly bounded with $|\Delta f_1(x_1, \eta_1, t)| \leq \rho_1(x_1)$, and $|\Delta f_2(x_1, x_2, \eta_2, t)| \leq \rho_2(x_1, x_2)$, for all η_1 and η_2 in some prescribed sets and for all (x_1, x_2, t) .
- A.3) For the functions g_i , there exist known continuous, positive definite functions β_i and nonnegative functions φ_i such that, for $i = 1, 2$ and for all $(x_1, x_2, u, \eta_1, \eta_2, t)$

$$\beta_1(|x_2|)(1 + \varphi_1(x_1, a_1 x_2, t)) \leq a_1 x_2 g_1(x_1, a_1 x_2, \eta_1, t),$$

$$\beta_2(|u|)(1 + \varphi_2(x_1, x_2, a_2 u, t)) \leq a_2 u g_2(x_1, x_2, a_2 u, \eta_2, t).$$

Moreover, there exist known continuous positive definite functions ϕ_i such that inequality $\beta_i(\phi_1(\phi_2)) \leq \phi_1 \phi_2 \phi_i$ holds for $i = 1, 2$ and for constants $\gamma_i \geq 0$ and $1 \leq \epsilon_i \leq 2$.

- A.4) The functions $g_i(\cdot)$ have the property that $g_1(x_1, a_1 x_2, \eta_1, t) = q_1(a_1)g_1(x_1, x_2, \eta_1, t)$ and $g_2(x_1, x_2, a_2 u, \eta_2, t) = q_2(a_2)g_2(x_1, x_2, u, \eta_2, t)$, where $q_1(\cdot)$ and $q_2(\cdot)$ are known, odd and normalized functions with $q_i(1) = 1$. The inverse functions of $q_i(\cdot)$ are assumed to be locally well defined. Moreover, $\partial g_1 / \partial x_2 \neq 0$ if $x_2 \neq 0$, and, if the sign of x_2 is fixed, the sign of the partial derivative is fixed as well for all x_1, η_1 and t .
- A.5) The functions in the system can be bounded by known functions that are uniformly bounded with respect to time and locally uniformly bounded with respect to the state. That is, for all $(x_1, x_2, u, \eta_1, \eta_2, t)$, $|f_1(x_1, t)| \leq \bar{f}_1(x_1)$, $|f_2(x_1, x_2, t)| \leq \bar{f}_2(x_1, x_2)$, $|g_1(x_1, a_1 x_2, \eta_1, t)| \leq \bar{g}_1(x_1, x_2)$, and $|g_2(x_1, x_2, a_2 u, \eta_2, t)| \leq \bar{g}_2(x_1, x_2, u)$. Also, the functions in the subsystem (1) are differentiable once and their partial derivatives are bounded by known functions as, for all $(x_1, x_2, u, \eta_1, \eta_2, t)$

$$\left| \frac{\partial f_1}{\partial t} \right| + \left| \frac{\partial \Delta f_1}{\partial t} \right| + \left| \frac{\partial g_1}{\partial t} \right| \leq \kappa_1(x_1, x_2),$$

$$\left| \frac{\partial f_1}{\partial x_1} \right| + \left| \frac{\partial \Delta f_1}{\partial x_1} \right| + \left| \frac{\partial g_1}{\partial x_1} \right| \leq \kappa_{11}(x_1, x_2, t),$$

$$\left| \frac{\partial g_1}{\partial x_2} \right| \leq \kappa_{12}(x_1, x_2, t).$$

It is worth mentioning that Assumptions A.1)–A.3) come directly from the generalized matching conditions [10]. Assumption A.4) is somewhat implied by Assumption A.3). Assumption A.5) is needed for identification of control directions. With these assumptions, we can proceed with continuous robust control design in the next section.

III. ROBUST CONTROL DESIGN

Along the line of nonlinear robust control design for uncertain systems, the main approach is Lyapunov's direct method together with using deterministic bounding functions for uncertainties. There are several classes of uncertain systems in [1], [3], [6], [9], [11] for which global stabilizing robust controllers have been found if bounding functions of uncertainties are available. One of the common features among these results is that stability analysis and especially generation of robust controller is done through a Lyapunov argument in which robust control is chosen to compensate for uncertainties by dominating them in both magnitude and sign. In those results, control directions are known in the sense that, after choosing robust controller, the sign of robust control determines the sign of the time derivative of the state and the Lyapunov function.

In this paper, robust design is studied for systems with unknown control direction. Like the existing results, the process of designing a successful robust control is to use the Lyapunov argument and the concept of control dominating unknown dynamics. Since control direction is unknown, an initial guess is made on control direction, and the robust control is designed whose direction is changed by a so-called shifting law. The key results in this study are that unknown control direction can be identified using the domination concept and that, since continuous control can produce a better transient response, smooth transition of the direction of robust control can be achieved. Intuitively, the basic idea emanates as following: begin with some guess a_1 of a_1 to design robust control in terms of \hat{a}_1 , and identify a_1 on line (both parts use the domination concept in a slightly different way); then formulate shifting laws \hat{a}_1 , so that robust control is continuous and guarantees global stability.

The following four-step procedure is used to design a continuous robust control without *a priori* knowledge of control direction: On-line identification of control directions, shifting laws, state transformation and robust control design using the recursive procedure in [10].

Step 1: The first step is to determine $\text{sign}[a_1]$ and $\text{sign}[a_2]$ on line. Without correct identification of these signs, any control will fail to ensure stability. As explained earlier, if a robust control works for a given uncertain system, the robust control must dominate the uncertainties in the system in both sign and magnitude. This domination provides us with a mean of determining the unknown control directions of the system.

Let us first investigate the subsystem (2). If the robust control u is designed properly, the function g_2 will dominate in magnitude and sign the uncertainty Δf_2 . Due to the unknown control direction a_2 , a successful robust control must identify the control direction, and the identification can be done by designing robust control such that the function g_2 dominates the control-unrelated part of dynamics, $f_2 + \Delta f_2$, that is, $|g_2(x_1, x_2, a_2 u, \eta_2, t)| \geq |f_2(x_1, x_2, t)| + |\Delta f_2(x_1, x_2, t)|$. This inequality can be guaranteed if u is chosen such that

$$g_2(x_1, x_2, |u|, \eta_2, t) \geq [\bar{f}_2(x_1, x_2) + \rho_2(x_1, x_2)]. \quad (3)$$

Note that, although the above inequality appears conservative for choosing robust control, the condition is necessary in the worst case that both the known and unknown dynamics f_2 and Δf_2 may be unstable and therefore, the control has to dominate.

Integrating both sides of (2) yields, for all $[t_1, t_2] \subset [t_0, t]$

$$x_2(t_2) - x_2(t_1) = \int_{t_1}^{t_2} [f_2(x_1, x_2, \tau) + \Delta f_2(x_1, x_2, \eta_2, \tau) + g_2(x_1, x_2, a_2 u, \eta_2, \tau)] d\tau.$$

Note that the integrand functions are all continuous. Under condition (3), it follows that, if $g_2(\cdot)$ does not change its sign for $\tau \in [t_1, t_2]$, $\text{sign}[x_2(t_2) - x_2(t_1)] = \text{sign}[g_2(x_1, x_2, a_2 u, \eta_2, \tau)]$. Using Assumptions A.3) and A.4) we have, for any $[t_1, t_2] \subset [t_0, t]$

$$\begin{aligned} \text{sign}[a_2] &= \text{sign}[g_2(a_2)] \\ &= \text{sign}[g_2(x_1, x_2, u, \eta_2, \tau)] \text{sign}[x_2(t_2) - x_2(t_1)] \\ &= \text{sign}[u(\tau)] \text{sign}[x_2(t_2) - x_2(t_1)] \quad \forall \tau \in [t_1, t_2] \end{aligned} \quad (4)$$

as long as condition (3) holds and $u(\tau)$ does not change its sign.

There are two important observations needed to be made about the identifications of $\text{sign}[a_2]$. First, since we can and have to choose robust control to dominate uncertainties to achieve stability, $\text{sign}[a_2]$ can be identified almost instantaneously after system (1) to (2) runs. Second, since a_2 is a scaled constant, its identification needs to be performed only once. That is, condition (3) is required to be true

only at one time instant. This makes it possible for us to smoothly transfer the robust control with potentially wrong control direction to that with correct control direction.

Applying the recursive design procedure introduced in [10], we can design robust control u through designing a fictitious control for subsystem (1), but the design of fictitious control requires identification of control direction $\text{sign}[a_1]$. Differentiating (1) yields

$$\begin{aligned} \ddot{x}_1 &= \left[\frac{\partial f_1}{\partial t} + \frac{\partial \Delta f_1}{\partial t} + \frac{\partial g_1}{\partial t} \right] + \left[\frac{\partial f_1}{\partial x_1} + \frac{\partial \Delta f_1}{\partial x_1} \right] \dot{x}_1 \\ &\quad + \left[\frac{\partial g_1}{\partial x_1} \dot{x}_1 + \frac{\partial g_1}{\partial x_2} \dot{x}_2 \right] \end{aligned}$$

which implies that, after substituting (1) and (2) into the above equation

$$\begin{aligned} \ddot{x}_1 &= h(x_1, x_2, \eta_1, \eta_2, t) + \frac{\partial g_1(x_1, a_1 x_2, \eta_1, t)}{\partial x_2} \\ &\quad \times g_2(x_1, x_2, a_2 u, \eta_2, t) \\ &= h(x_1, x_2, \eta_1, \eta_2, t) + q_1(a_1) q_2(a_2) \\ &\quad \times \frac{\partial g_1(x_1, x_2, \eta_1, t)}{\partial x_2} g_2(x_1, x_2, u, \eta_2, t) \end{aligned} \quad (5)$$

where function $h(\cdot)$ is defined and, by Assumption A.5), bounded as follows

$$\begin{aligned} &h(x_1, x_2, \eta_1, \eta_2, t) \\ &\triangleq \left[\frac{\partial f_1}{\partial t} + \frac{\partial \Delta f_1}{\partial t} + \frac{\partial g_1}{\partial t} \right] + \left[\frac{\partial f_1}{\partial x_1} + \frac{\partial \Delta f_1}{\partial x_1} + \frac{\partial g_1}{\partial x_1} \right] \\ &\quad \times (f_1 + \Delta f_1 + g_1) + \frac{\partial g_1}{\partial x_2} (f_2 + \Delta f_2). \\ &|h(x_1, x_2, \eta_1, \eta_2, t)| \\ &\leq h_1(x_1, x_2) + h_{x_1}(x_1, x_2) [\bar{f}_1(x_1) + \rho_1(x_1) + \bar{g}_1(x_1, x_2)] \\ &\quad + h_{x_2}(x_1, x_2) [\bar{f}_2(x_1, x_2) + \rho_2(x_1, x_2)] \triangleq \bar{h}(x_1, x_2). \end{aligned}$$

Similar to the process of determining $\text{sign}[a_2]$, we can find $\text{sign}[a_1]$ if robust control u is designed such that the sign of right hand side of (5) is decided through domination by the term containing $g_2(\cdot)$. That is, robust control is designed such that

$$\begin{aligned} &|g_2(x_1, x_2, a_2 u, \eta_2, t)| \\ &\geq \left[\frac{\partial g_1(x_1, a_1 x_2, \eta_1, t)}{\partial x_2} \right]^{-1} h(x_1, x_2, \eta_1, \eta_2, t) \end{aligned}$$

or more conservatively

$$g_2(x_1, x_2, |u|, \eta_2, t) \geq \left| \frac{\partial g_1}{\partial x_2} \right|^{-1} \bar{h}(x_1, x_2). \quad (6)$$

Recall from Assumptions A.3)–A.4) that $g_2(x_1, x_2, u, \eta_2, t)$ has the same sign as that of u . Under condition (6), we know from (5) that \ddot{x}_1 has the same sign as that of $q_1(a_1) q_2(a_2) u [\partial g_1(x_1, x_2, \eta_1, t) / \partial x_2]$. Then, over any time interval in which u and $\partial g_1 / \partial x_2$ do not change their sign, the sign of \ddot{x}_1 is fixed, which in turn implies that the solution x_1 has a fixed concavity or convexity. Hence, it follows from this geometric property that

$$\begin{aligned} \text{sign} \left[\frac{x_1(t_1) + x_1(t_2)}{2} - x_1 \left(\frac{t_1 + t_2}{2} \right) \right] &= \text{sign}[\ddot{x}_1(\tau)] \\ &\quad \forall t_1 \leq \tau \leq t_2. \end{aligned}$$

Thus, we have that, whenever no sign change occurs over the interval $t_1 \leq \tau \leq t_2$ for $u(\tau)$ and $\partial g_1(x_1, x_2, \eta_1, \tau)/\partial x_2$

$$\begin{aligned} \text{sign}[a_1] = & \text{sign}[a_2] \text{sign} \left[\frac{x_1(t_1) + x_1(t_2)}{2} - x_1 \left(\frac{t_1 + t_2}{2} \right) \right] \\ & \times \text{sign}[u(\tau)] \text{sign} \left[\frac{\partial g_1(x_1, x_2, \eta_1, \tau)}{\partial x_2} \right] \end{aligned} \quad (7)$$

provided that condition (6) holds. The Eqs. (7) and (4) show how to identify on line the control directions. Once the $\text{sign}[a_i]$ are determined, they can be used to construct the so-called shifting laws which direct continuously (fictitious) robust controls with guessed control directions to those with correct control directions. This will be done in the next step.

It may appear from (5) or (6) that there is a singularity problem when $|\partial g_1/\partial x_2|$ becomes arbitrarily small. Although this singularity problem does make it impossible for us to choose robust control u to satisfy (6) at the time instant that singularity occurs, it does not present any obstacle for identification of control direction a_1 . This is because condition (6) needs to hold only for a very small period of time and because singularity cannot present persistently since u chosen to dominate uncertainties and known dynamics will cause x_2 to increase in magnitude. This reasoning will become much more obvious after we state robust control laws and basic stability analysis. Thus for now, we proceed with our analysis by assuming

$$\left| \frac{\partial g_1}{\partial x_2} \right| > C_n \quad (8)$$

where C_n is a constant chosen by the designer.

Step 2: After both $\text{sign}[a_i]$ are determined, one has two options to implement robust control. The first option is to change the directions \hat{a}_i of robust controls instantaneously (at most once for each direction) to the correct values. If this option is adopted, the results in [10] can be readily applied to generate robust control and to conclude global stability. Since instantaneous changes of signs exist only in theory, sign changes should be implemented as fast as possible. Although this option is used sometimes in practice, it is not our choice because of the following problems. First, it is unclear how fast the change of control directions must be to ensure stability. Second, the rate of change of control directions is usually limited by control bandwidth and, even if possible, sign change being very fast is not desirable since it may excite high-frequency dynamics neglected in the course of modelling.

Our choice, the second option, is to transfer robust controls smoothly from possibly wrongly guessed control directions to the identified, correct control directions, which will be accomplished by formulating shifting laws. It will be shown that, as long as direction-independent robust control laws u_r and w_r are designed properly and implemented, global stability is guaranteed no matter what time constants and gains of the shifting laws are. This result provides a theoretical guarantee of the freedom for the designer to choose the gains in the shifting laws to achieve good transient performance. In fact, our choice gives a complete answer to the stability question posed for the first option. That is, in order not to cause saturation of actuators or to excite unmodeled dynamics while avoiding large overshoot of the state, shifting laws should be chosen such that their gains are reasonably large but not too large.

Let the initial guesses of the control directions a_1 and a_2 be denoted by $\hat{a}_1(t_0)$ and $\hat{a}_2(t_0)$, then the shifting laws are given by

$$\begin{aligned} \dot{c}_1 = & \begin{cases} 0 & \text{until condition} \\ -k_s c_1 - k_m \text{sign}[c_1] |\bar{g}_1(x_1, w_r)| & \text{(8) becomes valid} \\ & \text{thereafter} \end{cases} \quad (9) \\ \dot{c}_2 = & -k_s c_2 - k_m \text{sign}[c_2] |\bar{g}_2(x_1, x_2, w_r)| \quad (10) \end{aligned}$$

where $|\hat{a}_i(t_0)| = 1$, $c_i \triangleq \hat{b}_i - a_i$, $\hat{b}_i \triangleq g_i(\hat{a}_i)$, the variables \hat{a}_i and w_r will be chosen in steps three and four, $\bar{g}_1 \triangleq q_2^{-1}(\hat{b}_2) \cdot g_1^{-1}(\cdot)$ represents the inverse function of g_1 , $k_s, k_m > 0$ are the gains of the shifting laws, and $1/L$ is a positive constant. Both k_s and k_m can be freely chosen by the designer. The variables $\hat{a}_i(t)$ will be used as multipliers in robust control laws. The above shifting laws require identification results from (7) and (3). The initial values $|\hat{a}_i(t_0)| = 1$, as well as the part of setting $\hat{b}_1 = 0$ in (9), are chosen such that domination of robust controls with \hat{a}_i are in effect at least initially to perform identification of control directions. The choices of (9) and (10) are made so that, as will be shown in the proof, stability is warranted during the transition of control directions of robust controls.

In the subsequent design and analysis, the state will be shown to be uniformly bounded. More importantly, it is worth noting that, even though a large control of the wrong sign may be applied, the transient response of the state will not have a very large magnitude. This can be seen from the following facts. First, the singularity region is small in the state space. Second, whenever the state escapes from the singularity region, sign identification is almost instantaneous. Third, the shifting laws makes c_i converge to zero exponentially (and possibly in finite time); robust control will have the corrected sign whenever $|c_1| < 1$ and $|c_2| < 1$; and, in a finite time interval that can be made arbitrarily small by adjusting k_s , $|c_i|$ will be small enough so that robust control has its full strength. This implies that the temporary but adverse effect of large control of the wrong sign on the state can be minimized by choosing a reasonably large k_s .

Step 3: In this step, state transformation is performed to introduce a fictitious control variable w . This allows us to design robust control u recursively and to facilitate stability proof.

Let x_1^d be a bounded, smooth desired trajectory for the output x_1 of the system to track. Based on x_1^d and $x_1(0)$, we choose a smooth "perturbed desired" trajectory y_d to avoid singularity. The exact way of choosing y_d will be explained in step 4. Define the new state $z_1 = x_1 - y_d$, then (1) can be rewritten as

$$\begin{aligned} \dot{z}_1 = & f_1(x_1, t) + \Delta f_1(x_1, \eta_1, t) + g_1(x_1, a_1 x_2, \eta_1, t) - \dot{y}_d \\ = & -k z_1 + \Delta f_1'(z_1, x_1, \eta_1, t) + g_1(x_1, a_1 w, \eta_1, t) \\ & + [g_1(x_1, a_1 x_2, \eta_1, t) - g_1(x_1, a_1 w, \eta_1, t)] \end{aligned} \quad (11)$$

where $k > 0$ is a feedback gain chosen to be the designer, and the lumped uncertainty $\Delta f_1'(\cdot)$ is $\Delta f_1'(z_1, x_1, \eta_1, t) = f_1(x_1, t) + \Delta f_1(x_1, \eta_1, t) + k z_1 - \dot{y}_d$. Similar to classical observer design, k should be selected such that $k_s > k$ for good transient performance. It follows that

$$\begin{aligned} |\Delta f_1'(z_1, x_1, \eta_1, t)| & \leq \bar{f}_1(x_1) + \rho_1(x_1) + 0.5(1 + k^2 z_1^2) + |\dot{y}_d| \\ & \triangleq \rho_1'(z_1, x_1). \end{aligned}$$

It also follows that there exists a known function $\psi(\cdot)$ such that

$$|g_1(x_1, a_1 x_2, \eta_1, t) - g_1(x_1, a_1 w, \eta_1, t)| \leq |x_2 - w| \psi(x_1, x_2, w).$$

The choice of fictitious control w will be done in the next step. It will be shown that $w = w(x_1, z_1, \hat{a}_1)$. Now, define the second new state $z_2 = x_2 - w$. Then, we can rewrite (2) as

$$\begin{aligned} \dot{z}_2 = & f_2(x_1, x_2, t) + \Delta f_2(x_1, x_2, \eta_2, t) \\ & - \dot{w} + g_2(x_1, x_2, a_2 u, \eta_2, t) \\ \triangleq & -k z_2 - \psi(x_1, x_2, w) |z_1| \text{sign}[z_2] \\ & + \Delta f_2'(z_1, x_1, z_2, x_2, w, \eta_2, t) \\ & + g_2(x_1, x_2, a_2 u, \eta_2, t) \end{aligned} \quad (12)$$

where the "lumped uncertainty" is $\Delta f_2'(z_1, x_1, z_2, x_2, w, \eta_2, t) = f_2(x_1, x_2, t) + \Delta f_2(x_1, x_2, \eta_2, t) + k z_2 + \psi(x_1, x_2, w) |z_1| \text{sign}[z_2]$

— \dot{u}_2 . The uncertainty can be bounded by a known function $\rho'_2(\cdot)$ as

$$\begin{aligned} & |\Delta f'_2(z_1, x_1, z_2, x_2, w, \eta_2, t)| \\ & \leq \bar{f}_2(x_1, x_2) + \rho_2(x_1, x_2) \\ & + 0.5(1 + k^2 z_2^2) + \psi(x_1, x_2, w) |z_1| \\ & + \left| \frac{\partial w}{\partial \hat{a}_1} \right| \cdot |\dot{\hat{a}}_1| + \left| \frac{\partial w}{\partial z_1} \right| \cdot |z_1| \\ & + \left| \frac{\partial w}{\partial x_1} \right| \cdot |\dot{x}_1| \leq \rho'_2(z_1, x_1, z_2, x_2) \end{aligned} \quad (13)$$

where $|z_1|$, $|\dot{x}_1|$ and $|\dot{\hat{a}}_1|$ can be easily bounded, and the partial derivatives of w with respect to its arguments can be bounded as well after w is explicitly designed in the next step. It has been shown in [10] that bounding functions and robust control can be made differentiable. Therefore, the bounding function $\rho'_2(\cdot)$ in (13) can be found.

Step 4: In this step, robust control u and fictitious control w are designed to guarantee global stability. Based on stability analysis presented in the appendix, they are chosen respectively to be of the form

$$w(z_1, x_1, \hat{a}_1) = -\hat{a}_1 \varrho_1(z_1, x_1) \phi_1(\rho''_1(z_1, x_1)) \triangleq -\hat{a}_1 w_1 \quad (14)$$

$$\begin{aligned} u(z_1, x_1, z_2, x_2, \hat{a}_2) &= -\hat{a}_2 \varrho_2(z_1, x_1, z_2, x_2) \\ &\quad \times \phi_2(\rho''_2(z_1, x_1, z_2, x_2)) \\ &\triangleq -\hat{a}_2 u_1 \end{aligned} \quad (15)$$

where

$$\begin{aligned} \varrho_1(z_1, x_1) &\triangleq \frac{\mu_1^2(z_1, x_1) + \epsilon^2}{|\mu_1(z_1, x_1)|^3 + \epsilon^3} \mu_1(z_1, x_1), \\ \mu_1(z_1, x_1) &= z_1 \rho''_1(z_1, x_1) \\ \varrho_2(z_1, x_1, z_2, x_2) &\triangleq \frac{\mu_2^2(z_1, x_1, z_2, x_2) + \epsilon^2}{|\mu_2(z_1, x_1, z_2, x_2)|^3 + \epsilon^3} \\ &\quad \times \mu_2(z_1, x_1, z_2, x_2), \\ \mu_2(z_1, x_1, z_2, x_2) &= z_2 \rho''_2(z_1, x_1, z_2, x_2) \end{aligned}$$

where $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are functions defined in Assumption A.3), $\epsilon > 0$ is a design parameter. Note that $|\varrho_i| \leq 2$ and, if $|\mu_i| \geq \epsilon$, $|\varrho_i| \geq 1$. The bounding functions $\rho''_1(\cdot)$ and $\rho''_2(\cdot)$ should be chosen to be continuous and to satisfy the following conditions

$$\rho''_1(z_1, x_1) \geq \rho'_1(z_1, x_1), \quad \rho''_1(z_1, x_1)|_{t=t_0} \neq 0,$$

$$\rho''_2(z_1, x_1)|_{t=t_0} \neq 0,$$

$$\rho''_2(z_1, x_1, z_2, x_2) \geq \max \left\{ \rho'_2(z_1, x_1, z_2, x_2), \left[\bar{f}_2(x_1, x_2) + \rho_2(x_1, x_2) \right], \frac{\bar{h}(x_1, x_2)}{C_a} \right\}.$$

Implementation of the above robust controllers requires identification of control directions by domination. The domination can be ensured by choosing y_d and ϵ as follows. First, select $y_d(t_0)$ such that $z_2(t_0) \neq 0$ (which can be done by adjusting $z_1(t_0)$ through choosing y_d). Then, it follows that, at time t_0 , robust control u is nonzero. Note that $|\hat{a}_2(t_0)| = 1$. Also note that ϵ can be chosen freely and that, as will be shown later, the smaller ϵ the better the tracking performance. Thus, we can choose ϵ such that, at time $t = t_0$, $|\mu_2(z_1, x_1, z_2, x_2)| > \epsilon$, which implies that $1 \leq |\varrho_2(z_1, x_1, z_2, x_2)| \leq 2$ at time t_0 . Therefore, we know from Assumption A.3) that

$$\begin{aligned} g_2(x_1, x_2, |u|, \eta_2, t) &\geq \frac{\beta_2(|u|)}{|u|} = \frac{\beta_2(\varrho_2 \phi_2(\rho''_2))}{|u|} \\ &\geq \frac{\varrho_2 \phi_2(\rho''_2) \rho''_2}{|u|} = \rho''_2(z_1, x_1, z_2, x_2) \end{aligned}$$

which, based on the choice of $\rho''_2(\cdot)$, guarantees both inequalities (6) and (3). And these two inequalities can always be satisfied for any finite interval if ϵ is small and if $y_d(t)$ is perturbed from $x'_d(t)$ such that $z_2(t)$ is not zero in that interval. This concludes the four-step process of designing smooth robust control law.

Global stability of the system under robust control laws and shifting laws are analyzed by Lyapunov's direct method using Lyapunov function $L(z_1, z_2, \hat{a}_1, \hat{a}_2) = L_1(z_1, \hat{a}_1) + L_2(z_2, \hat{a}_2)$ where

$$L_1(z_1, \hat{a}_1) = z_1^2 + z_2^2,$$

$$L_2(z_2, \hat{a}_2) = \frac{1}{k_m} (a_1 q_1(\hat{a}_1) - 1)^2 + \frac{1}{k_m} (a_2 q_2(\hat{a}_2) - 1)^2. \quad (16)$$

The following theorem guarantees global uniform ultimate boundedness of the system in the form of (11) to (12). The proof can be found in the Appendix.

Theorem. Suppose that condition (8) is valid. Then, under the robust control laws (14) and (15) together with the shifting laws (9) and (10), the state of system (11) to (12) is globally uniformly ultimately bounded. Moreover, the system output x_1 tracks any smooth, bounded desired trajectory $x'_d(t)$ with error that can be made arbitrarily small in the sense that

$$\limsup_{t \rightarrow \infty} |x_1(\tau) - x'_d(\tau)| \leq \sqrt{\frac{2\epsilon k_m + \delta}{k k_m}}.$$

Although the above theorem excludes the singularity problem in its statement, one can extend the result of the theorem to a more general case through the following argument. First, nonsingularity condition (8) needs to be valid only for one instant or a small period of time. Second, C_a is a design parameter and can be made smaller to make singularity region in the state space smaller. Third, since uncertainties as well as known dynamics are bounded by well defined functions, \dot{x}_1 is finite at time t_0 for any initial conditions. Thus, when the system does start in the singularity region, robust control u dominates all dynamics and therefore, by choosing a small enough C_a , can make $|x_2|$ large and out of the singularity region quickly enough to identify the control directions. One may argue that this scheme does not work when subsystem (1) has arbitrarily small finite escape time if the state starts in the singularity region. However, the case does not exist since it contradicts \dot{x}_1 being finite. And, if this case would exist, there is no control, continuous or discontinuous, of finite magnitude that can stabilize the whole system.

Illustrative Example

Consider the following second-order nonlinear system

$$\dot{x}_1 = a_1^1 x_2^1 + \Delta f, \quad \dot{x}_2 = a_2 x_1 u$$

where the uncertainty $\Delta f = x_1^2 \cos(t) + 0.5 \sin(0.5t)$. Robust controls (14) and (15) are chosen with the design parameters $\epsilon = 0.1$, $k = 2$, $C_a = 0.2$, and with the following bounding functions

$$\rho''_1 = \rho'_1 = x_1^2 + 0.5 + \frac{k}{2}(1 + z_1^2),$$

$$\rho''_2 = \max \left\{ \rho'_2, \frac{\bar{h}}{3x_2^2} \right\},$$

$$\bar{h} = 0.25 + 2|x_1|(x_1^2 + 0.25 + |x_2|^4),$$

$$\begin{aligned} \rho'_2 &= k|z_2| + |z_1|z_2^2 + 12 \frac{\mu_1^2 + \epsilon_1^2}{|\mu_1|^3 + \epsilon_1^3} \rho_1'^2 [0.25 + x_1^2 + |a_1|^4 |x_2|^4] \\ &\quad + \left[\frac{\mu_1^2 + \epsilon_1^2}{|\mu_1|^3 + \epsilon_1^3} |z_1| \rho_1'^2 \right]^4 |z_1| \\ &\quad + \frac{k}{|a_1|^4} |a_1^1 \hat{a}_1^1 - 1| \left[\frac{\mu_1^2 + \epsilon_1^2}{|\mu_1|^3 + \epsilon_1^3} |z_1| \rho_1'^2 \right]. \end{aligned}$$

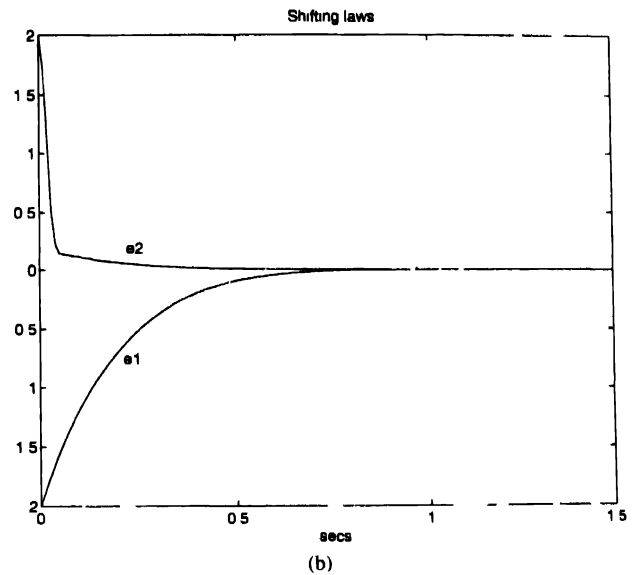
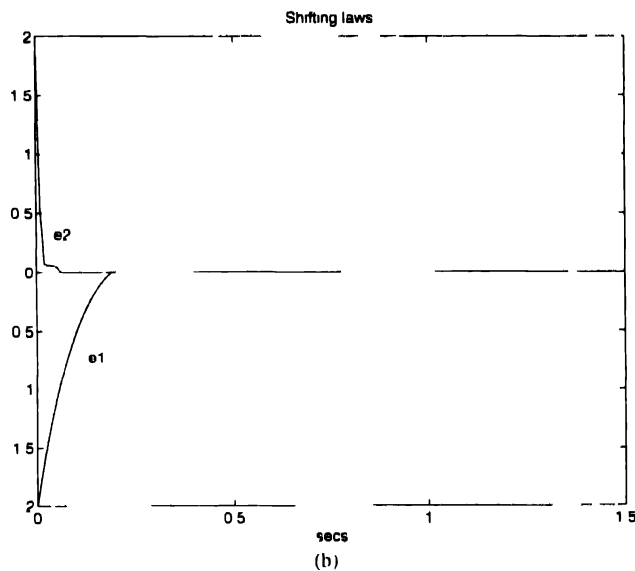
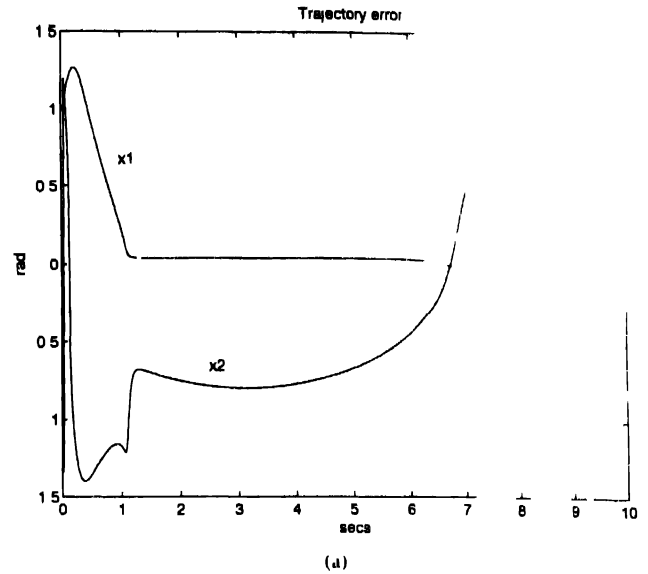
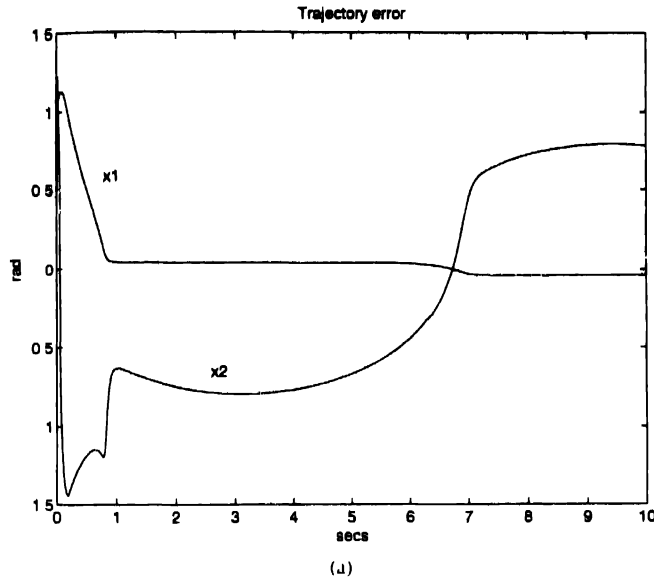


Fig 1 Robust control with $k = 10$ and $k = 1$

Fig 2 Robust control with $k = 5$ and $k = 0.1$

The simulation was carried out using SIMMON[©] with the following choices (1) desired trajectory is $y^T = x^T = 0$ (2) initial conditions are $x_1(0) = x_2(0) = 1$ $a_1(t_0) = -1$ $a_2(t_0) = 1$ (3) unknown control directions are chosen to be $a_1 = 1$ and $a_2 = -1$

The simulation results of the states and the shifting laws are shown in Figs. 1 and 2 respectively for two sets of choices k and k . It should be noted that there is no large overshoot in the state during and after sign identification.

IV. CONCLUSION

Robust output tracking control of a class of nonlinear uncertain systems is studied. A system in the class satisfies the generalized matching conditions but contains one or several parameters of unknown signs used to characterize control directions. Without *a priori* knowledge of control directions, a four step design procedure is proposed which consists of identification of control directions and construction of robust controls and shifting laws. The system may contain nonlinear uncertainties, and both continuous and discontinuous robust control can be designed using only bounding functions

of unknown dynamics. Under the condition that the subsystem(s) not directly controlled has no arbitrarily small finite escape time, the proposed robust control guarantees global uniform ultimate bounded stability with arbitrarily small ultimate output tracking error.

APPENDIX

Proof of the Theorem Taking the time derivative of $I_1(t)$ in (16) along the trajectory of system (11) and (12) under control (14) and (15) we get

$$\begin{aligned} \dot{I}_1 = & 2 \{ -k_1 + \Delta f'_1(x_1, x_2, u, \eta, t) + g_1(x_1, x_2, u, \eta, t) \\ & + [q_1(x_1, x_2, u, \eta, t) - q_1(x_1, x_2, u, \eta, t)] \} \\ & + 2 \{ -k_2 - \epsilon(x_1, x_2, u) | \text{sign}[\cdot] | \\ & + \Delta f'_2(x_1, x_2, u, \eta, t) + q_2(x_1, x_2, u, \eta, t) \} \\ \leq & -2k_1 L_1(x_1, x_2) + 2|\mu_1(x_1, x_2)| + 2|\mu_2(x_1, x_2)| \\ & + 2|q_1(x_1, x_2, u, \eta, t)| + 2|q_2(x_1, x_2, u, \eta, t)| \\ & + 2|-\epsilon| |q_1(x_1, x_2, u, \eta, t)| + 2|-\epsilon| |q_2(x_1, x_2, u, \eta, t)| \end{aligned}$$

Using the same arguments in [10], one can show by Assumption A.3) that $z_i g_i(\cdot) \leq 0$ and that, if $|\mu_i(\cdot)| > \epsilon$, $z_i g_i(\cdot) \leq -|\mu_i(\cdot)|$ for $i = 1, 2$. Therefore, we have that, for all (z_i, x_i, t)

$$\begin{aligned} \dot{L}_1 \leq & -2kL_1 + 4\epsilon + 2|z_1|c_1|\bar{g}_1(x_1, w_1) \\ & + 2|z_2|c_2|\bar{g}_2(x_1, x_2, u_1). \end{aligned} \quad (17)$$

Now we proceed by taking the time derivative of L_2 in (16) along shifting law (9) and (10). It follows from $\text{sign}[g_i(a_i)] = q_i(a_i) = a_i$ that

$$\begin{aligned} \dot{L}_2(\hat{a}_1, \hat{a}_2) = & \frac{2}{k_m} a_1(a_1 q_1(\hat{a}_1) - 1)\dot{r}_1 + \frac{2}{k_m} a_2(a_2 q_2(\hat{a}_2) - 1)\dot{r}_2 \\ \leq & -2\frac{k_s}{k_m} L_2(\hat{a}_1, \hat{a}_2) - 2|z_1|c_1|\bar{g}_1(x_1, w_1) \\ & - 2|z_2|c_2|\bar{g}_2(x_1, x_2, u_1). \end{aligned}$$

Combining the above inequality with (17) we have that, for all (x_1, z_1, x_2, z_2, t) , $\dot{L}_1 + \dot{L}_2 \leq -2kL_1 - 2k_s L_2/k_m + 4\epsilon$. The above inequality shows that the state of system (11) to (12) is globally uniformly ultimately bounded. Solving the inequality and noting that $|\hat{a}_i(0)| = 1$, we can show the ultimate bound on the output tracking error in the statement of the theorem.

REFERENCES

- [1] M. J. Corless and G. Leitmann, "Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamical systems," *IEEE Trans. Automat. Contr.*, vol. 26, pp. 1139-1144, 1981.
- [2] F. Giri, P. A. Ioannou, and Ahmed-Zaid, "A stable adaptive control scheme for first order plants with no priori knowledge on the parameters," *IEEE Trans. Automat. Contr.*, vol. 38, no. 5, pp. 766-771, May 1993.
- [3] S. Gutman, "Uncertain dynamical systems—A Lyapunov min max approach," *IEEE Trans. Automat. Contr.*, vol. AC-24, no. 3, June 1979.
- [4] R. Lozano and B. Brogliato, "Adaptive control of a simple nonlinear system without *a priori* information on the plant parameter," *IEEE Trans. Automat. Contr.*, vol. 37, no. 1, pp. 30-37, Jan. 1992.
- [5] R. Lozano, J. Collado, and S. Mondie, "Model reference adaptive control without *a priori* knowledge of the high frequency gain," *IEEE Trans. Automat. Contr.*, vol. 35, no. 1, pp. 71-78, Jan. 1990.
- [6] R. Marino and P. Tomei, "Robust stabilization of feedback linearizable time-varying uncertain nonlinear systems," *Automatica*, vol. 29, pp. 181-189, 1993.
- [7] D. R. Mudgett and A. S. Morse, "Adaptive stabilization of linear systems with unknown high frequency gain," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 549-554, June 1985.
- [8] R. D. Nussbaum, "Some remarks on the conjecture in parameter adaptive control," *Syst. Contr. Lett.*, vol. 3, no. 5, pp. 243-246, 1983.
- [9] Z. Qu, "Global stabilization of nonlinear systems with a class of unmatched uncertainties," *Sys. Contr. Lett.*, vol. 18, no. 3, pp. 301-307, 1992.
- [10] —, "Robust control of nonlinear uncertain systems under generalized matching conditions," *Automatica*, vol. 29, pp. 985-998, July 1993.
- [11] J. J. E. Slotine and K. Hedrick, "Robust input-output feedback linearization," *Int. J. Contr.*, vol. 57, pp. 1133-1139, 1993.

On the Ordering of Optimal Hedging Points in a Class of Manufacturing Flow Control Models

George Liberopoulos and Jian-Qiang Hu

Abstract—The optimal flow control policy of a single-product unreliable manufacturing system that must meet a constant demand rate is known to be a threshold type policy: safety production surplus levels called hedging points (thresholds) are associated with each discrete stochastic capacity state of the system and serve to protect the production process from uncertainty in future capacity availability. This correspondence extends and generalizes previous results on the ordering of optimal hedging points. Our method is based on examining special properties of the Bellman optimality conditions of the underlying stochastic control problem.

1. INTRODUCTION

Manufacturing flow control models address the dynamic allocation of stochastic capacity among competing products in a just in time manufacturing environment. Several of these models assume that medium-term demand rates for a number of products are constant, while short-term production rates are continuous over time and must live within capacity constraints. These constraints are dictated by the configuration of operational machines in the manufacturing system modeled. When the state of machines (e.g., operational/failed) changes at random points in time, optimal policies are characterized by generally intractable dynamic programming Bellman equations [1]. To date, most analytical manufacturing flow control models have assumed that the machine state process is homogeneous and uncontrolled, implying that the failure rate of a machine does not depend on its production rate or its age, i.e., how long that machine has been operational since its most recent failure.

Akella and Kumar [2] were the first to perform rigorous analysis on a one-product, one-machine (with two machine states, up and down) manufacturing system with a discounted cost criterion. For the same system, but with a long-run average cost criterion, Bielecki and Kumar [4] derived the steady state probability distribution of production surplus. Sharifnia [17] extended this approach to one-product, multiple-machine-state systems, and Algoet [3] derived partial differential equations for the joint steady state probability density function of production surplus for multiple-product systems, but did not solve these equations. Caramanis and Sharifnia [5] used the results of [17] to design near optimal control policies for multiple-product systems by decomposing them to many analytically tractable one-product systems. Malhamé and Boukas [13] and Malhamé [14] demonstrated the Markov renewal nature of one-product manufacturing systems under hedging point policies and studied the ergodicity of such policies. Liberopoulos and Caramanis [12], Hu *et al.* [7], and Yu and Vakili [19] explored the optimal control structure for one-product systems with production dependent failure rates. Hu and Xiang [8] demonstrated the equivalence of a one-product, two-machine-state manufacturing system with a queueing system. Hu and Xiang [9], [10] showed that the ordering of optimal hedging points of certain one-product systems can be found analytically. Sethi *et al.* [16] analyzed the structure of turnpike sets that characterize optimal production

Manuscript received December 20, 1993; revised March 30, 1994. This was supported in part by the National Science Foundation under Grants EID-9212122 and DDM-9212368.

The authors are with the Department of Manufacturing Engineering, College of Engineering, Boston University, Boston, MA 02215 USA.
IEEE Log Number 9406986.

surplus levels when the demand is a Markov chain; they showed that the turnpike sets exhibit a monotone property with respect to capacity and demand. A more comprehensive list of references on manufacturing flow control may be found in a recently published book by Gershwin [6].

A key feature of the optimal control structure common to all models is the hedging point. A hedging point is a production surplus threshold level associated with each discrete machine state i , toward which the production surplus must be guided as quickly as possible while the machine state remains fixed at i . A hedging point is also referred to as a safety or buffer production surplus and is carried to protect the production process from the uncertainty in future capacity availability or other uncertainties (e.g., demand and processing time).

This correspondence mainly deals with systems similar to those studied in [14], [17], having one product with constant demand rate, and many machine states with constant transition rates. It is known that apart from a one-product, two-machine-state system [2], [4], it is practically impossible to derive explicit expressions for the optimal hedging points. With this in mind, the goal of this correspondence is to uncover properties on the relative ordering of optimal hedging points. Such properties have many important applications. They can be used to develop simple numerical and heuristic methods for obtaining optimal or near-optimal control policies, since they dramatically reduce the search space of optimal hedging points. Also, for systems where machine state transition rates are age dependent (age being the time elapsed since the most recent machine state change), they can be used to study properties of the corresponding optimal age dependent hedging curves or turnpikes. This can be done by approximating age dependent transition time distributions with phase type distributions, each phase having associated with it a hedging point. This correspondence extends and generalizes results reported by Hu and Xiang [9], [10], who showed that for some special systems optimal hedging points are ordered monotonically. The method we use to establish the relative ordering of optimal hedging points is based on the Hamilton–Jacobi–Bellman optimality conditions of the underlying stochastic control problem.

The remaining of the correspondence is organized as follows. In Section II the manufacturing flow control problem is formulated as an optimal control problem whose optimality conditions are given by a set of Hamilton–Jacobi–Bellman equations. In Section III certain properties on the ordering of optimal hedging points are derived based on ensuring consistency of signs on both sides of the derivative of the Hamilton–Jacobi–Bellman equation. In Section IV these properties are applied to several special systems, and in Section V several issues and extensions are discussed.

II. THE MANUFACTURING FLOW CONTROL MODEL

We consider a manufacturing system with the following features [17]:

- 1) The system produces a single part-type for which there is a constant demand rate denoted d .
- 2) The capacity state (also called machine state) of the system is an finite-state continuous-time irreducible Markov chain denoted $\alpha(t)$, with state space S .
- 3) The production rate of the system at time t , denoted $u(t)$, must belong to the closed interval $U_{\alpha(t)} = \{u: 0 \leq u \leq r_{\alpha(t)}\}$, where r_i is the maximum production rate of the system at machine state $i \in S$.
- 4) The production surplus of the system at time t , denoted $x(t)$, is the difference between the cumulative production and cumulative demand up to time t .
- 5) The production surplus cost rate, denoted $g(x)$, is a strictly convex function in x and has a unique minimum at x^* which implies

that $g(x) \rightarrow \infty$ as $x \rightarrow \pm\infty$; $g(x)$ represents the (per unit, backlog) cost when x is positive (respectively, backlog).

- 6) The transition time from machine state i to machine state j is exponentially distributed with constant transition rate $q(i, j)$, $i, j \in S$, with $q(i, i) = -\sum_{j \neq i} q(i, j) \in \mathbb{N}$, $0, i \neq j$ and $q(i, i) < 0, i \in S$.

The last feature implies that the machine state process $\{\alpha(t), t \geq 0\}$ is homogeneous and uncontrolled. With this in mind, $\{x(t), \alpha(t), t \geq 0\}$ is a stochastic process characterized by the following differential equation:

$$dx(t)/dt = u(t) - d, \quad \text{s.t. } 0 \leq u(t) \leq r_{\alpha(t)} \quad (1)$$

The objective is to find optimal production rates $u_i^*(x) \in U_i$ to minimize the long-run expected average cost

$$J = \lim_{t \rightarrow \infty} E \left\{ (1/t) \int_0^t g(x(s)) ds \right\}. \quad (2)$$

For the above limit to exist it is assumed that in the long-run the demand is strictly feasible, i.e., that the demand is less than the equilibrium mean of the capacity process. Under suitable regularity conditions imposed on the control [15],¹ the optimal feedback control of the system is characterized by the following Hamilton–Jacobi–Bellman (HJB) dynamic programming equation [6]:

$$J^* = g(x) + \min_{u \in U_{\alpha(x)}} \{V'_{\alpha(x)}(x)(u - d)\} + \sum_{i \neq \alpha(x)} q(i, \alpha(x)) [V_i(x) - V_{\alpha(x)}], \quad x \in S, \quad (3)$$

where $J^* = \inf_{\{u_i(\cdot)\}} J$, $V_i(x)$ is the optimal value or differential cost-to-go function starting from state (i, x) , and $'$ denotes the derivative w.r.t. x . According to (3), the optimal production rate $u_i^*(x)$ satisfies $u_i^*(x) = \arg \min \{V'_{\alpha(x)}(x)u: 0 \leq u \leq r_i\}$.

Theorem 1: $V_i(x)$ is strictly convex in x and has a unique minimum at $x_i^*, i \in S$.

Proof Outline The proof that $V_i(x)$ is strictly convex is similar to that of Theorem 5.1 in [18] and we shall omit it. We only note, however, that the strict convexity of $V_i(x)$ follows from the strict convexity of $g(x)$ and the convexity of the sets $U_i, i \in S$. In fact, exact solution of $V_i(x)$ for small systems [2], [4], [16] suggests that the value function may be strictly convex even if $g(x)$ is not strictly convex. For the results obtained in [18] to hold, it must be assumed that under any admissible control law some state (x_0, k) is a regeneration point of the process $\{x(t), \alpha(t)\}$ (for example k may be any machine state and x_0 any point such that $x_0 \leq x^*$). Denote by T_n successive visit (regeneration) times to (x_0, k) . The value $V_i(x)$ is well defined for regenerative control laws as

$$V_i(x) = \min_{\{u_i(\cdot)\}} E \left\{ \int_0^{T_1} (g(x(s)) - J) ds \mid x(0) = x, \alpha(0) = i \right\}. \quad (4)$$

That $V_i(x)$ has a unique minimum then follows from its strict convexity and the fact that $g(x) \rightarrow \infty$ as $x \rightarrow \pm\infty$, since, by (4) this fact implies that $V_i(x) \rightarrow \infty$ as $x \rightarrow \pm\infty$. Q.E.D.

Theorem 1 implies that

$$\begin{aligned} V'_i(x) &> 0, & V''_i(x) &> 0 & \text{for } x > x_i^*, \\ V'_i(x) &< 0, & V''_i(x) &> 0 & \text{for } x < x_i^*. \end{aligned} \quad (5)$$

In view of the above, the optimal control policy is given by,

$$u_i^*(x) = \begin{cases} r_i & \text{if } x < x_i^*, \\ \min(d, r_i) & \text{if } x = x_i^*, \\ 0 & \text{if } x > x_i^*. \end{cases} \quad (6)$$

¹As we consider the infinite horizon case we also require that admissible control laws be restricted to those for which the process $\{x(t), \alpha(t)\}$ is asymptotically stationary ergodic.

Point z_i is known as the optimal hedging point associated with machine state i . Unfortunately, the value of the $z_i, i \in S$, can not be explicitly calculated except for a system with one machine that dwells in two states (up and down) [2], [4]. For systems with more than two machine states, however, the optimal hedging points can be obtained numerically: first, the long-run average cost can be computed for each tentative choice of hedging points, $z_i, i \in S$, as follows:

$$J^* = E[g(x)] = \int_{-\infty}^{+\infty} g(x)f^*(x)dx,$$

where $f^*(x)$ is the steady-state probability density function of x and is readily available as an implicit function of the $z_i, i \in S$ [17]. Then, the optimal values of the z_i can be found numerically by minimizing J^* using a gradient search algorithm.

III. HEDGING POINT ORDERING PROPERTIES

Recognizing that it is practically impossible to derive explicit expressions for the optimal hedging points, we instead seek to uncover properties of the relative ordering of hedging points. To this end, we assume that the function $V_i(x), i \in S$, exist and are continuously differentiable,² and that the optimal control is characterized by (3). If $V_i(x), i \in S$, is continuously differentiable, $V_i''(x)$ is defined almost everywhere (except possibly at z_i and the points where $g'(x)$ is not defined). Differentiating (3) w.r.t. x yields

$$0 = g'(x) + V_i''(x)(u_i^*(x) - d) + \sum_{j \neq i} q(i, j)[V_j'(x) - V_i'(x)], \quad \text{a.e. } i \in S. \quad (7)$$

The proofs of the properties that follow are based on ensuring consistency of signs of quantities appearing on both sides of (7). Namely, the right-hand side of (7) cannot be strictly positive (or negative), since the left-hand side is 0.

In what follows it is assumed that $r_i \neq d$, for all $i \in S$. This assumption simplifies matters but is not essential. Under this assumption, there are two types of machine states, feasible and infeasible. State i is called feasible if $r_i > d$, and infeasible if $r_i < d$.

Moreover, it is assumed that there is at least one infeasible machine state. Otherwise, d can be met at all times, and there is no need to hedge beyond x^* , the minimizer of $g(x)$. If all states are feasible, all hedging points are equal to x^* .

Property 1: $x^* \leq \min_{k \in S} \{z_k\}$.

Proof: Suppose the opposite is true, i.e., suppose that $x^* > z_i = \min_{k \in S} \{z_k\}$. Then, for $x = z_i^+$, the first two terms on the right-hand side of (7) are strictly negative, the terms in the summation are nonpositive, whereas the left-hand side is 0. Therefore, $x^* > \min_{k \in S} \{z_k\}$ cannot be true. Q.E.D.

In most manufacturing systems $x^* = 0$, i.e., the cost rate is minimum when there is neither positive (inventory) nor negative (backlog) production surplus. In these cases, Property 1 states that safety production surplus levels must always be nonnegative.

In most of the work reported to date [17], [13], [14] it has been assumed, without proof, that optimal hedging points associated with infeasible machine states are larger than the largest of the hedging points associated with feasible machine states. This means that in infeasible machine states it is best to always produce at the maximum allowable rate. Although intuitively this seems reasonable, it has yet to be proved; however, we can show the following properties.

²For $V_i(x)$ to be continuously differentiable, one must further assume that $g(x)$ is Lipschitz continuous. Continuous differentiability of the $V_i(x)$ has been proved for special manufacturing flow systems [2], [4] with piece-wise constant cost rate function, as well as general piece-wise deterministic systems with jump Markov disturbances [15]; it has also been proved for systems with random demand [16].

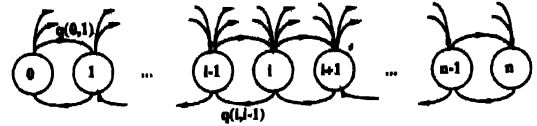


Fig. 1. System 1.

Property 2: If $r_i > d$, for some $i \in S$, then $z_i \leq \max_{j: q(i, j) > 0} \{z_j\}$. Equivalently, if $z_i > \max_{j: q(i, j) > 0} \{z_j\}$, for some $i \in S$, then $r_i < d$.

Proof: For the first part, suppose the opposite is true, i.e., suppose that $z_i > \max_{j: q(i, j) > 0} \{z_j\}$. Then, for $x = z_i^+$, the first two terms on the right-hand side of (7) are strictly positive, the terms in the summation are nonnegative, whereas the left-hand side is 0. Therefore, $z_i > \max_{j: q(i, j) > 0} \{z_j\}$ cannot be true. The second part is dual to the first part and follows immediately from it. Q.E.D.

It is noteworthy that the inequality $z_i \leq \max_{j: q(i, j) > 0} \{z_j\}$ (respectively, $z_i > \max_{j: q(i, j) > 0} \{z_j\}$) in Property 2 becomes, $z_i < \max_{j: q(i, j) > 0} \{z_j\}$ (respectively, $z_i \geq \max_{j: q(i, j) > 0} \{z_j\}$), if either $z_i > x^*$ or $g(x)$ is continuously differentiable. A corollary of Property 2 is that the largest hedging point belongs to an infeasible machine state. This is stated as Property 3.

Property 3: $\max_{i: r_i < d} \{z_i\} \leq \max_{i: r_i > d} \{z_i\}$.

The following property concerns the ordering of any two hedging points.

Property 4: For $i \neq j, i, j \in S$, define

$$\bar{z} = \max_{k: q(i, k) \geq q(j, k), k \neq i, j} \{z_k\} \\ (\bar{z} = -\infty, \text{ if } \{k: q(i, k) \geq q(j, k), k \neq i, j\} = \emptyset), \\ \underline{z} = \min_{k: q(i, k) \leq q(j, k), k \neq i, j} \{z_k\} \\ (\underline{z} = \infty, \text{ if } \{k: q(i, k) \leq q(j, k), k \neq i, j\} = \emptyset).$$

If $r_i > d, z_i \geq \bar{z}, z_j \leq \underline{z}$, and $\bar{z} \leq \underline{z}$, then $z_i \leq z_j$.

Proof: Suppose the opposite is true, i.e., suppose that $z_i > z_j$. Subtract (7) for machine state j from that for machine state i . After rearranging terms, the result is

$$0 = V_i''(x)(u_i^*(x) - d) - V_j''(x)(u_j^*(x) - d) \\ + \sum_{k \neq i, j} [q(i, k) - q(j, k)]V_k'(x) \\ - q(i, k)V_i'(x) + q(j, k)V_j'(x) \\ + [q(i, j) + q(j, i)][V_j'(x) - V_i'(x)].$$

Then, for $x: \max\{\bar{z}, z_j\} < x < \min\{z_i, \underline{z}\}$, all but the terms in the summation on the right-hand side of the above equation are strictly positive, the terms in the summation are nonnegative, whereas the left-hand side is 0. Therefore, $z_i > z_j$ cannot be true. Q.E.D.

There may be little intuition about Property 4 at this point; more intuition about this property may be gained when it is seen in the context of the special system examples in Section IV. It should be pointed out that Properties 1–3 hold even if the $V_i(x), i \in S$ are not strictly convex.

IV. SPECIAL SYSTEM EXAMPLES

Hu and Xiang [9], [10] studied optimal hedging point ordering properties for systems with special Markov chain structure under a discounted cost criterion. In this section we demonstrate how optimal hedging point ordering properties of those and other systems are special cases of the general properties developed in Section III, for the average cost criterion. All the systems we study have only one infeasible machine state, labeled 0.

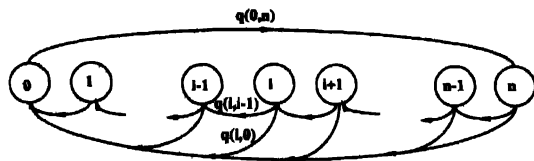


Fig. 2 System 2

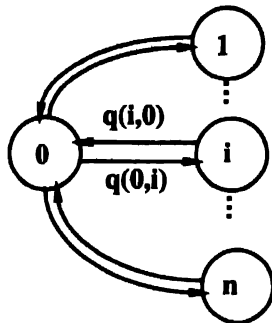


Fig. 3 System 3

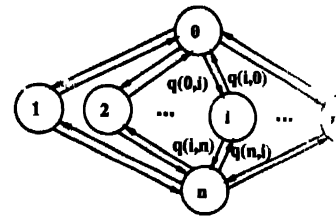


Fig. 4 System 4

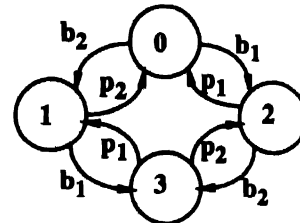


Fig. 5 A system with two nonidentical unreliable machines. There are four possible machine states (0, 0), (0, 1), (1, 0), (1, 1) where a 1 (respectively 0) in position i indicates that the i th machine is up (respectively down). These states are denoted 0, 1, 2, 3 respectively. The failure and repair rates of the i th machine are p_i and b_i respectively.

System 1 is shown in Fig. 1 and is defined as follows.

System 1 Definition $S = \{0, 1, \dots, n\}$ where $r_0 < d$ and $r_i > d$, $i = 1, 2, \dots, n$. The following transitions *must* be allowed: $i \rightarrow i-1$, $i = 1, \dots, n$ and $i \rightarrow n$ for some $i < n$. The *only other* allowable transitions are $i \rightarrow j$, $j = i+1, \dots, n-1$.

The characteristic of System 1's Markov chain is that any transition from left to right is allowed, whereas transitions from right to left must be *skip free*. A special instance of System 1 is when transitions are skip free from left to right too (death-birth chain). A death-birth chain can be used to model a manufacturing system with n identical machines whose machine state is characterized by the number of machines that are up. System 1 has the following property.

Property 5 For System 1, $r^* < r$, $r_i \leq r$, $i = 1, 2, \dots, n$.

Proof By Property 1, $r^* \leq r$, $r_i = 0$, $i = 1, 2, \dots, n$. The rest of the proof is by induction. First by Property 2, $r_1 \leq r$. Now suppose that $r_i \leq r$, $i = 1, 2, \dots, i-1$, $0 < i < n$. We must prove that $r_i \leq r$. Suppose that the opposite is true, i.e., suppose that $r_i > r$. This implies that $r_i > \max\{r_{i-1}, r_{i+1}\} = \max\{r_{i-1}, r_{i+1}\}$. Consequently, by Property 2, $r_i < d$. However, $r_i < d$ is not true; therefore, $r_i > r$ cannot be true either. Instead, $r_i \leq r$ must be true. Q.E.D.

System 2 is shown in Fig. 2 and is defined as follows.

System 2 Definition $S = \{0, 1, \dots, n\}$ where $r_0 < d$ and $r_i > d$, $i = 1, 2, \dots, n$. The following transitions *must* be allowed: $i \rightarrow i-1$, $i = 1, 2, \dots, n$ and $0 \rightarrow n$. No other transitions are allowed.

A special instance of System 2 is when $r_0 = 0$ and $r_i = r$, $i = 1, 2, \dots, n$. This can be used to model a system with one machine that may be either up or down and whose down time is exponentially distributed and up time has a Coxian distribution with n stages. A special case of the latter system is a system where the only transition allowed to machine state 0 is from machine state 1. Such a system can be used to model an Erlang up time distribution with n phases. System 2 has the following property.

Property 6 For System 2,

- 1) If $q(n, 0) \leq q(n-1, 0) \leq \dots \leq q(1, 0)$ then $r^* \leq r$, $r_i \leq r$, $i = 1, 2, \dots, n$.
- 2) If $q(n, 0) \geq q(n-1, 0) \geq \dots \geq q(1, 0)$ then $r^* \leq r$, $r_i \leq r$, $i = 1, 2, \dots, n$.

- 3) If $q(n, 0) = q(n-1, 0) = \dots = q(1, 0)$ then $r^* = r$, $r_i = r$, $i = 1, 2, \dots, n$.

Proof We only prove 1) since the proof for 2) is very similar and 3) follows immediately from 1) and 2). By Property 1, $r^* \leq r$, $r_i = 0$, $i = 1, 2, \dots, n$. The rest of the proof is by induction. First by Property 2, $r_1 \leq r$. Now suppose that $r_i \leq r$, $i = 1, 2, \dots, i-1$, $0 < i < n$. We must prove that $r_i \leq r$. That $r_i \leq r$ follows immediately from Property 4 after substituting r_{i-1} by r_{i+1} , $r_i = r$ respectively and noting that $q(i+1, 0) \leq q(i, 0)$. Q.E.D.

System 3 is shown in Fig. 3 and is defined as follows.

System 3 Definition $S = \{0, 1, \dots, n\}$ where $r_0 < r$ and $r_i > d$, $i = 1, 2, \dots, n$. The following transitions *must* be allowed: $i \leftrightarrow 0$, $i = 1, 2, \dots, n$. No other transitions are allowed.

When $r_0 = 0$ and $r_i = r$, $i = 1, 2, \dots, n$, System 3 can be used to model a system with one machine that may be either up or down and whose down time is exponentially distributed and up time has a hyperexponential distribution with n states. System 3 has the following property.

Property 7 For System 3, $r^* \leq r$, $r_i = r$, $i = 1, 2, \dots, n$. Also, if $q(i, 0) \leq q(j, 0)$, $(i, j) \in \{1, 2, \dots, n\}$ then

Proof By Property 1, $r^* \leq r$, $r_0 = 0$, $r_i = r$, $i = 1, 2, \dots, n$. The last part follows immediately from Property 4 after substituting r_{i-1} by r_{j-1} respectively and noting that $q(i, 0) \leq q(j, 0)$. Q.E.D.

System 4 is shown in Fig. 4 and is defined as follows.

System 4 Definition $S = \{0, 1, \dots, n\}$ where $r_0 < d$ and $r_i > d$, $i = 1, 2, \dots, n$. The following transitions *must* be allowed: $i \leftrightarrow 0$ and $i \leftrightarrow n$, $i = 1, 2, \dots, n-1$. No other transitions are allowed.

System 4 can be used to model a production system that may enter one of many failure modes before failing completely. For example, a system with two nonidentical unreliable machines is a special case of System 4 with $n = 3$ (see Fig. 5). For System 4 we have the following result.

Property 8 For System 4,

- 1) $r^* < r \leq \max_{k=1, \dots, n} \{r_k\}$.
- 2) If there exists $j^* \in S$ such that $q(j^*, 0) = \max_{k=1, \dots, n} \{q(k, 0)\}$ and $q(j^*, n) = \min_{k=1, \dots, n} \{q(k, n)\}$ then $r_j = \max_{k=1, \dots, n} \{r_k\}$.

Proof:

1) It follows immediately from Properties 1, 2, and 3.
 2) Suppose the opposite is true, i.e., suppose that $z_{i^*} = \max_{k: 1 \leq k \leq n-1} \{z_k\}$, where $i^* \neq j^*$. Since $q(i^*, n) \geq q(j^*, n)$, $q(i^*, 0) \leq q(j^*, 0)$, and $x^* \leq z_n \leq z_{i^*} \leq z_0$, following the proof of Property 4, it can be easily shown that $z_{i^*} > z_{j^*}$ cannot be true; hence $z_{i^*} \leq z_{j^*}$ must be true. This implies that $z_{j^*} = \max_{k: 1 \leq k \leq n-1} \{z_k\}$. Q.E.D.

V. FURTHER ISSUES

In Sections III and IV it was seen that for a system whose machine state process is modeled as a homogeneous and uncontrolled Markov chain, the relative ordering of optimal hedging points can be determined, if this Markov chain has a special structure and, in some cases, its transition rates have a certain ordering. It is noteworthy that this ordering of hedging points does not depend on the specific values of r_i and $q(i, j)$, $i, j \in S$. A question that arises is the following. Once the ordering of hedging points has been established using the type of arguments presented in Section III, can similar arguments be used to also determine the ordering of the differences between consecutive hedging points? Based on numerical examples, our conjecture is that the ordering of the differences between consecutive hedging points depends on the specific values of r_i and $q(i, j)$, $i, j \in S$.

An extension of the manufacturing flow control model presented in Section II is a model with age-dependent machine state process. By age dependent we mean that transition rates of the machine state process depend on the age of the machine state, i.e., on how long the machine state process has been in a particular state. We first note that for a system with age dependent machine state process the optimal control may depend not only on the machine state of the system but also on the age of the machine state. Consequently, the optimal control is no longer a hedging point policy. Instead, it is a so-called hedging curve or turnpike policy—a generalized hedging point policy in which values of hedging points depend on the age of the machine state.

In [10] it is shown that any transition (failure) time with age dependent transition (failure) rate can be approximated arbitrarily closely (in the sense of weak convergence) by a random variable with Coxian³ distribution with many stages. The stages of the Coxian distribution correspond to different ages of the transition (failure) time, and the transition (failure) rates between the stages are proportional to the corresponding values of the hazard rate of the original age dependent transition (failure) time distribution. Since a random variable with Coxian distribution is a Markov process itself, a system with age dependent machine state process can be approximated by a system with homogeneous Markov machine state. This approximation provides a means to investigate ordering properties of the optimal control policy for a system with age dependent machine state process based on results obtained on a related system with homogeneous Markov machine state process.

Finally, an extension of systems with homogeneous Markov machine state processes is systems in which transition rates of machine states depend on machine production rates. Some results on such systems have been reported in [7], [12], [19].

REFERENCES

- [1] S. Asmussen, *Applied Probability and Queues*. New York: Wiley, 1987.
- [2] R. Akella and P. R. Kumar, "Optimal control of production rate in a failure prone manufacturing system," *IEEE Trans. Automat. Contr.*, vol. 31, pp. 116–126, 1986.

³For a definition of the Coxian distribution, see, for example, [1].

- [3] P. H. Algoet, "Flow balance equations for the steady-state distribution of a flexible manufacturing system," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 917–921, 1989.
- [4] T. Bielecki and P. R. Kumar, "Optimality of zero-inventory policies for unreliable manufacturing systems," *Oper. Res.*, vol. 36, pp. 532–541, 1988.
- [5] M. Caramanis and A. Sharifnia, "Near-optimal manufacturing flow controller design," *Int. J. Flex. Manuf. Syst.*, vol. 3, pp. 321–336, 1991.
- [6] S. B. Gershwin, *Manufacturing Systems Engineering*. Englewood Cliffs, NJ: Prentice-Hall, 1994.
- [7] J.-Q. Hu, P. Vakili, and G.-X. Yu, "Optimality of hedging point policies in the production control of failure prone manufacturing systems," *IEEE Trans. Automat. Contr.*, vol. 39, no. 9, pp. 1875–1880, 1994.
- [8] J.-Q. Hu and D. Xiang, "The queueing equivalence to a manufacturing system with failures," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 499–502, 1993.
- [9] —, "Structural properties of optimal controllers in failure prone manufacturing systems," *IEEE Trans. Automat. Contr.*, vol. 39, no. 3, pp. 640–643, 1994.
- [10] —, "Monotonicity of optimal controls for unreliable manufacturing systems," to appear in *J. Opt. Theory Applications*.
- [11] J. Kimemia and S. B. Gershwin, "An algorithm for the computer control of production in flexible manufacturing systems," *IEE Trans.*, vol. 15, pp. 353–362, 1983.
- [12] G. Liberopoulos and M. Caramanis, "Production control of manufacturing systems with production rate dependent failure rates," *IEEE Trans. Automat. Contr.*, vol. 39, no. 4, 1994.
- [13] R. P. Malhamé and E.-K. Boukas, "A renewal theoretic analysis of a class of manufacturing systems," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 580–587, 1991.
- [14] R. P. Malhamé, "Ergodicity of hedging control policies in single-part multiple-state manufacturing systems," *IEEE Trans. Automat. Contr.*, vol. 38, no. 2, pp. 340–343, 1993.
- [15] R. Rishel, "Dynamic programming and minimum principles for systems with jump Markov disturbances," *SIAM J. Contr.*, vol. 13, pp. 338–371, 1975.
- [16] S. Sethi, H. M. Soner, Q. Zhang, and J. Jiang, "Turnpike sets and their analysis in stochastic production planning problems," *Math. OR*, vol. 17, pp. 932–950, 1992.
- [17] A. Sharifnia, "Production control of a manufacturing system with multiple machine states," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 620–625, 1988.
- [18] J. N. Tsitsiklis, "Convexity and characterization of optimal policies in a dynamic routing problem," *J. Opt. Theory Applications*, vol. 44, pp. 105–136, 1984.
- [19] G. X. Yu and P. Vakili, "Control of manufacturing systems with production dependent machine failures," in *Proc. 31st Annu. Allerton Conf. Communication, Control, and Computing*, Allerton, IL, 1993.

Recursive Identification Method for MISO Wiener-Hammerstein Model

M. Boutayeb and M. Darouach

Abstract—A simple technique for recursive identification of the Wiener-Hammerstein model with extension to the multi-input single-output (MISO) case is presented. We use a new transformation of the input-output difference equation where parameters to be estimated are those of each subsystem of the initial and unique realization. After that, a weighted extended least squares (WELS) method is employed to estimate recursively and separately parameters of the linear subsystems and the static nonlinear element. Convergence analysis of the proposed procedure is also studied. Finally, a numerical example is provided to show the efficiency of the algorithm.

I. INTRODUCTION

Modeling, identification, and control design of nonlinear systems have been the subject of many research activities in the last decades. Indeed, for many dynamic systems the use of nonlinear models is often of great interest and generally characterizes adequately physical processes over their whole operating range [8]. Thus, accuracy and performances of the control law increase significantly. One of the nonlinear realizations frequently studied is the Hammerstein model, which is composed of a static nonlinearity in series with a linear dynamic system. Several identification techniques of this kind of models with an extension to the MISO model were performed. The literature is abundant about this subject and we refer the reader to [1], [9], and the references therein. However, all recursive identification methods of SISO and MISO Hammerstein model ([5], [6], and [11]) consist at first in transforming the initial realization to an equivalent representation with common denominator and extended numerators. Next to that, classical recursive methods are applied. The obtained parameter vector to be estimated by this approach, is of high dimension, particularly for large scale systems, and then computational requirements increase with possible numerical instabilities. Thus, control design becomes very complicated and too crude to be of use.

On the other hand, very few efforts have been done to extend these methods to the general representation, called the Wiener-Hammerstein model or G-model, particularly the MISO case. The G-model is defined as a linear system in cascade with a static nonlinear element followed by another linear system. In [1] and [2], Billings *et al.* have proposed an identification algorithm for the Wiener-Hammerstein model based on correlation analysis. This was earlier suggested by Korenberg [10]. However, in their algorithms, some restrictive assumptions are required for the kind of input sequences to preserve the separability principle. On the other hand, computational requirements are considerable. More recently, Yoshine *et al.* [18] have suggested another approach for identification of the G-model, this consists of estimating, in the SISO case, impulse responses of the linear subsystems and parameters of the nonlinear element.

In this paper, we present a recursive method to estimate parameters of the linear and nonlinear parts of the G-model, which is an extension

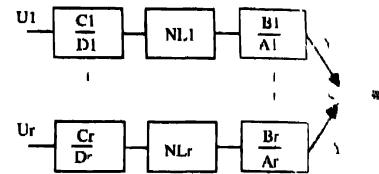


Fig. 1 MISO Wiener-Hammerstein model

of our previous work in [5]. We investigate a new formulation of the input-output difference equation, from which a WELS algorithm is established. Contrary to the approach developed in [6] and [11], the obtained algorithm has the advantage to estimate separately and recursively parameters of the linear and nonlinear subsystems of the initial realization where parameter number is of minimal dimension. All these results are extended to the MISO Wiener-Hammerstein model. Conditions for parameter convergence to the actual ones are established, it's also shown that the weighted factors are introduced to enhance the convergence of the proposed algorithm.

II. PROBLEM FORMULATION

Fig. 1 shows the general scheme of the Wiener-Hammerstein discrete time model with r inputs. The output signal y_k at time k is given by

$$y_k = y_{1k} + \dots + y_{rk} \quad (1)$$

where y_{rk} is the nonmeasured intermediate output related to the r th input

$$y_{rk} = \frac{B_r(q^{-1})}{A_r(q^{-1})} v_{rk} \quad (2)$$

with

$$v_{rk} = q_{r1} Z_k + \dots + q_{rn} Z_k \quad (3)$$

and

$$Z_k = \frac{C_r(q^{-1})}{D_r(q^{-1})} u_{rk} \quad (4)$$

The linear subsystems in the G-model are defined as

$$A_i(q^{-1}) = 1 + a_{i1}q^{-1} + \dots + a_{in_i}q^{-n_i} \quad (5)$$

$$B_i(q^{-1}) = b_{i0} + b_{i1}q^{-1} + \dots + b_{ip_i}q^{-p_i} \quad (6)$$

$$C_i(q^{-1}) = c_{i0} + c_{i1}q^{-1} + \dots + c_{is_i}q^{-s_i} \quad (7)$$

and

$$D_i(q^{-1}) = 1 + d_{i1}q^{-1} + \dots + d_{if_i}q^{-f_i} \quad \text{for } i = 1, \dots, r \quad (8)$$

q^{-1} is the delay operator, u_{rk} is the r th input of the system at time instant k and orders f_i, m_i, n_i, p_i and s_i are supposed to be known. These can be estimated from the input-output data [9]. For simplicity in the sequel, we note $A_i(q^{-1})$, $B_i(q^{-1})$, $C_i(q^{-1})$ and $D_i(q^{-1})$ by A_i , B_i , C_i and D_i , respectively, $i = 1, \dots, r$. The linear transfer functions $B_i(q^{-1})/A_i(q^{-1})$ and $C_i(q^{-1})/D_i(q^{-1})$ are assumed to be stable.

III. SOME PRELIMINARY TRANSFORMATIONS

Before giving the main results of this paper, let us make some important transformations which lead to a unique and equivalent realization of (1)–(9).

Manuscript received July 8, 1993; revised December 21, 1993 and April 11, 1994.

The authors are with CRAN-FARAI-CNRS UA 821, Université de Nancy I, 186, Rue de Lorraine, 54400 Cosnes et Romain, France.
IEEE Log Number 9406987.

We remark that there are an infinity of realizations, noted by $(A_i, B_i, C_i, D_i, g_{ij})$, equivalent to $(A_i, B_i, C_i, D_i, g_{ij})$ for $i = 1, \dots, r$ and $j = 1, \dots, m_i$. Indeed, the input output representation of the G-model may be written as

$$y_k = y_{1k} + \dots + y_{rk} \quad (10)$$

with

$$y_{ik} = \frac{B_i}{A_i} V_{ik} \quad (11)$$

$$V_{ik} = g_{i1} Z_{ik} + \dots + g_{im_i} Z_{ik}^{m_i} \quad (12)$$

$$Z_{ik} = \frac{C_i}{D_i} u_{ik} \quad (13)$$

$$B_i = b_{i0} + b_{i1}q^{-1} + \dots + b_{ip_i}q^{-p_i} \quad (14)$$

and

$$C_i = c_{i0} + c_{i1}q^{-1} + \dots + c_{is_i}q^{-s_i} \quad \text{for } i = 1, \dots, r \quad (15)$$

the obtained parameters are related to the previous ones by

$$b_{ij} = l_i b_{i,j}, \quad c_{ij} = t_i c_{i,j} \quad \text{and} \quad g_{ij} = \frac{g_{i,j}}{l_i t_i} \quad (16)$$

for any l_i and $t_i \in \mathbb{R}^*$. Polynomials A_i and D_i are those in (6) and (9). One interesting way to choose l_i and t_i is to set

$$l_i = b_{i0} \quad \text{and} \quad t_i = c_{i0} \quad (17)$$

polynomials B_i and C_i are then in the following form

$$B_i(q^{-1}) = 1 + b_{i1}q^{-1} + \dots + b_{ip_i}q^{-p_i} \quad (18)$$

and

$$C_i(q^{-1}) = 1 + c_{i1}q^{-1} + \dots + c_{is_i}q^{-s_i} \quad \text{for } i = 1, \dots, r. \quad (19)$$

The final realization, where the index i is omitted, is summarized as follows:

$$y_k = y_{1k} + \dots + y_{rk} \quad (20)$$

with

$$y_{ik} = \frac{B_i}{A_i} V_{ik} \quad (21)$$

$$V_{ik} = g_{i1} Z_{ik} + \dots + g_{im_i} Z_{ik}^{m_i} \quad (22)$$

$$Z_{ik} = \frac{C_i}{D_i} u_{ik} \quad (23)$$

$$B_i = 1 + b_{i1}q^{-1} + \dots + b_{ip_i}q^{-p_i} \quad (24)$$

and

$$C_i = 1 + c_{i1}q^{-1} + \dots + c_{is_i}q^{-s_i} \quad \text{for } i = 1, \dots, r. \quad (25)$$

We remark that the transformation in (16), (17) allows us to replace b_{i0} and c_{i0} by 1, then parameters number to be estimated is n_I instead of $(nr + 2r)$ in the initial realization. On the other hand we can show easily that (20)–(25) is unique, this property assures the uniqueness of the identified parameters. Order n_I is defined as

$$n_I = f + n + m + s + p$$

with $f = f_1 + \dots + f_r$, $n = n_1 + \dots + n_r$, $m = m_1 + \dots + m_r$, $s = s_1 + \dots + s_r$ and $p = p_1 + \dots + p_r$.

IV. RECURSIVE PARAMETER ESTIMATION OF SISO WIENER-HAMMERSTEIN MODEL

For the case of SISO Wiener-Hammerstein model, the output signal is

$$y_k = \frac{B}{A} \left(g_1 \frac{C}{D} u_k + \dots + g_m \left(\frac{C}{D} u_k \right)^m \right) \quad (26)$$

with

$$A = 1 + a_1 q^{-1} + \dots + a_n q^{-n} \quad (27)$$

$$B = 1 + b_1 q^{-1} + \dots + b_p q^{-p} \quad (28)$$

$$C = 1 + c_1 q^{-1} + \dots + c_s q^{-s} \quad (29)$$

and

$$D = 1 + d_1 q^{-1} + \dots + d_l q^{-l}. \quad (30)$$

Into the difference form and by making use of (26) the output signal may be written explicitly as

$$y_k = F(\theta, y_k, u_k, Z_k). \quad (31)$$

θ contains parameters to be estimated of the realization (26)–(30) and is defined as

$$\theta = (a_1 \dots a_n \quad b_1 \dots b_p \quad c_1 \dots c_s \quad d_1 \dots d_l \quad g_1 \dots g_m)^T \quad (32)$$

$$F(\theta, y_k, u_k) = -\bar{A}y_k + g_1 B Z_k + \dots + g_m B Z_k^m \quad (33)$$

$$Z_k = C u_k - \bar{D} Z_k \quad (34)$$

and

$$\bar{A} = a_1 q^{-1} + \dots + a_n q^{-n}, \quad \bar{D} = d_1 q^{-1} + \dots + d_l q^{-l}. \quad (35)$$

We note that Z_k was omitted on the left-hand side of (33). Indeed, computation of Z_k is recursively done by (34) and will depend only on the input signal for all initial values Z_0, \dots, Z_l (because the transfer function C/D is assumed to be stable).

Recursive parameter estimation of the Wiener-Hammerstein model is given by the following theorem.

Theorem 1. A WERLS estimation of the parameter vector θ of the Wiener-Hammerstein model is

$$\theta_{k+1} = \theta_k + K_{k+1}(y_{k+1} - F(\theta_k, y_{k+1}, u_{k+1}))$$

$$K_{k+1} = P_k f_{k+1} (f_{k+1}^T P_k f_{k+1} + \lambda_{k+1})^{-1}$$

$$P_{k+1} = (I - K_{k+1} f_{k+1}^T) P_k$$

with

$$f_{k+1} = \frac{\partial F(\theta, y_{k+1}, u_{k+1})}{\partial \theta} \bigg|_{\theta = \theta_k}$$

$\{\lambda_{k+1}\}$ is a sequence of weighted factors that are positive numbers.

Proof. The WERLS estimation of θ is obtained by minimizing the classical criterion function

$$J_{k+1} = \frac{1}{k+1} \sum_{j=1}^{k+1} \frac{1}{\lambda_j} (y_j - \hat{y}_j)^2 \quad (36)$$

where \hat{y}_j can be seen as a prediction of y_j and is defined as

$$\hat{y}_j = F(\hat{\theta}_k, y_j, u_j) \quad (37)$$

where $\hat{\theta}_k$ represents the parameter vector estimation of θ by minimizing J_k .

$\hat{\theta}_{k+1}$ is then obtained by setting the derivative of J_{k+1} , with respect to θ , equal to zero:

$$\frac{\partial J_{k+1}}{\partial \theta} = 0 \Leftrightarrow \sum_{j=1}^{k+1} \frac{1}{\lambda_j} \frac{\partial y_j}{\partial \theta} (y_j - \hat{y}_j) = 0 \quad (38)$$

$$\Leftrightarrow \hat{\theta}_{k+1} = \left[\sum_{j=1}^{k+1} \frac{1}{\lambda_j} f_j f_j^T \right]^{-1} \sum_{j=1}^{k+1} \frac{1}{\lambda_j} f_j y_j \quad (39)$$

where $\partial y_j / \partial \theta$ is approximated by

$$f_j = \left| \frac{\partial F(\theta, y_j, u_j)}{\partial \theta} \right|_{\theta = \theta_k} \quad (40)$$

Then we can prove easily that in the recursive form [14], we obtain

$$\theta_{k+1} = \theta_k + \Gamma_{k+1}(y_{k+1} - F(\theta_k, y_{k+1}, u_{k+1})) \quad (41)$$

$$\Gamma_{k+1} = P_k f_{k+1} (f_{k+1}^T P_k f_{k+1} + \lambda_{k+1})^{-1} \quad (42)$$

$$P_{k+1} = (I - \Gamma_{k+1} f_{k+1}^T) P_k \quad (43)$$

and

$$f_{k+1} = \left. \frac{\partial F(\theta, y_{k+1}, u_{k+1})}{\partial \theta} \right|_{\theta=\theta_k} \quad (44)$$

End of proof.

V RECURSIVE PARAMETER ESTIMATION OF MISO WIEBER HAMMERSTEIN MODEL

In this section a direct extension of the proposed method in the SISO case is presented. At first we introduce the signal vector as in (31) where θ contains parameters of each sub system of the realization (20)–(25) next we use the same principle as in the SISO case

From (20) the output signal may be written explicitly as

$$y_k = \frac{B_1}{A_1} \left(q_{11} \frac{C_1}{D_1} u_{k+1} + \dots + q_{1, l-1} \left(\frac{C_1}{D_1} u_{1k} \right)^{l-1} \right) + \frac{B}{A} \left(q_{11} \frac{C}{D} u_{k+1} + \dots + q_{1, l-1} \left(\frac{C}{D} u_{1k} \right)^{l-1} \right) \quad (45)$$

or

$$\prod_{i=1}^l A_i y_k = \sum_{i=1}^l F_i (q_{11} Z_k + \dots + q_{1, l-1} Z_k^{l-1}) \quad (46)$$

with

$$F_i = B_i \prod_{j=1}^l A_j \quad \text{and} \quad Z_k = C u_k - \bar{D} Z_k \quad (47)$$

As $A_i = 1$ for $i = 1$ contains the term $\prod_{i=1}^l A_i$ is always in the form

$$\prod_{i=1}^l A_i = 1 + \bar{A} \quad \text{and} \quad D = d_{11} q^{-1} + \dots + d_{1, l-1} q^{l-1} \quad (48)$$

\bar{A} is a polynomial nonlinear in parameters with a known structure

From (48) and by making use of (46) y_k becomes

$$y_k = F(\theta, y_k, u_k) \quad (49)$$

with

$$\theta = (a_{11}, a_{1, l-1}, a_{1, l-2}, \dots, a_{1, 1}, b_{11}, b_{1, l-1}, b_{1, l-2}, \dots, b_{1, 1}, c_{11}, c_{1, l-1}, c_{1, l-2}, \dots, c_{1, 1}, d_{11}, d_{1, l-1}, d_{1, l-2}, \dots, d_{1, 1}, q_{11}, q_{1, l-1}, q_{1, l-2}, \dots, q_{1, 1})^T \quad (50)$$

and

$$F(\theta, y_k, u_k) = -\bar{A} y_k + \sum_{i=1}^l F_i (q_{11} Z_k + \dots + q_{1, l-1} Z_k^{l-1}) \quad (51)$$

thus we obtain the same formulation as in the nonlinear SISO case with multiple inputs where θ contains parameters of each subsystem of the initial realization. The WERLS estimator of θ is then given by Theorem 1

VI CONVERGENCE ANALYSIS

Several approaches have been proposed for the parameter convergence analysis of RLS algorithms in the deterministic case. Here we extend the method developed in [19] to the general case of polynomial realization (49)

Theorem 2 If the assumptions

$$\lim_{k \rightarrow \infty} \lambda_{\min} [P_k^{-1}] = \infty$$

$$\limsup_{k \rightarrow \infty} \frac{\lambda_{\max} [P_k^{-1}]}{\lambda_{\min} [P_k^{-1}]} < \infty$$

$$1 - \sqrt{\Delta_{k+1}} < \alpha_{k+1} < 1 + \sqrt{\Delta_{k+1}}$$

hold then the estimator given by Theorem 1 ensures that

$$\lim_{k \rightarrow \infty} \theta_k = \theta \quad (55)$$

where

$$\Delta_{k+1} = 1 - \frac{f_{k+1}^T P_k f_{k+1}}{f_{k+1}^T P_k f_{k+1} + \lambda_{k+1}} \quad (56)$$

and $\lambda_{\min} [P_k^{-1}]$, $\lambda_{\max} [P_k^{-1}]$ are the minimum and maximum eigenvalues of $[P_k^{-1}]$ respectively. α_{k+1} will be determined later.

Proof Consider a quadratic function defined as

$$V_{k+1} = \theta_{k+1}^T P_{k+1}^{-1} \theta_{k+1} \quad (57)$$

with

$$\theta_{k+1} = \theta_k - \theta \quad (58)$$

By subtracting θ from both sides of (41) we obtain

$$\theta_{k+1} = \theta_k + \Gamma_{k+1} f_{k+1} (f_{k+1}^T P_k f_{k+1} + \lambda_{k+1})^{-1} \epsilon_{k+1} \quad (59)$$

with

$$\epsilon_{k+1} = y_{k+1} - F(\theta_k, y_{k+1}, u_{k+1}) \quad (60)$$

On the other hand from (42)–(43) we have

$$P_{k+1} f_{k+1} = \lambda_{k+1} P_k f_{k+1} (f_{k+1}^T P_k f_{k+1} + \lambda_{k+1})^{-1} \quad (61)$$

Substituting (61) into (59) and (59) into (57) respectively the quadratic function becomes

$$V_{k+1} = \left[\theta_k + \frac{1}{\lambda_{k+1}} P_{k+1} f_{k+1} \epsilon_{k+1} \right]^T P_{k+1}^{-1} \left[\theta_k + \frac{1}{\lambda_{k+1}} P_{k+1} f_{k+1} \epsilon_{k+1} \right] \quad (62)$$

$$\Leftrightarrow V_{k+1} = \theta_k^T P_{k+1}^{-1} \theta_k + \frac{2}{\lambda_{k+1}} \theta_k^T f_{k+1} \epsilon_{k+1} + \frac{1}{\lambda_{k+1}} f_{k+1}^T P_{k+1}^{-1} f_{k+1} \epsilon_{k+1}^2 \quad (63)$$

From (42)–(43) we have

$$P_{k+1}^{-1} = P_k^{-1} + f_{k+1} \frac{1}{\lambda_{k+1}} f_{k+1}^T \quad (64)$$

Then (63) becomes

$$V_{k+1} = \theta_k^T P_k^{-1} \theta_k + \theta_k^T f_{k+1} \frac{1}{\lambda_{k+1}} f_{k+1}^T \theta_k + \frac{2}{\lambda_{k+1}} \theta_k^T f_{k+1} \epsilon_{k+1} + \frac{1}{\lambda_{k+1}} f_{k+1}^T P_k^{-1} f_{k+1} \epsilon_{k+1}^2 \quad (65)$$

As $f_{k+1}^T \theta_k$ is an approximation of the innovation sequence ϵ_{k+1} we introduce an unknown factor α_{k+1} to correct this approximation such as

$$f_{k+1}^T \theta_k = -\alpha_{k+1} \epsilon_{k+1} \quad (66)$$

for linear systems and in [19] we have $f_k = \Phi_k$ which is independent from θ and then $\alpha_{k+1} = 1$. Substituting (66) in (65) we obtain

$$V_{k+1} = V_k + \left(\frac{\alpha_{k+1}^2 - 2\alpha_{k+1}}{\lambda_{k+1}} + \frac{1}{\lambda_{k+1}} f_{k+1}^T P_k^{-1} f_{k+1} \right) \epsilon_{k+1}^2 \quad (67)$$

thus, if the condition (54) holds, we have

$$\frac{\alpha_{k+1}^2 - 2\alpha_{k+1}}{\lambda_{k+1}} + \frac{1}{\lambda_{k+1}} \hat{f}_{k+1}^T P_{k+1} \hat{f}_{k+1} < 0 \quad (68)$$

and then

$$V_{k+1} \leq V_k \leq \dots \leq V_0. \quad (69)$$

We note

$$\lim_{k \rightarrow \infty} V_k = V.$$

From the persistently exciting conditions (52) and (53), we have

$$\frac{V_k}{\text{tr}(P_k^{-1})} \geq \frac{\lambda_{\min}[P_k^{-1}] \theta_k^T \theta_k}{n_I \lambda_{\max}[P_k^{-1}]} \geq 0 \quad (70)$$

and then

$$\lim_{k \rightarrow \infty} \frac{V_k}{\text{tr}(P_k^{-1})} = 0 \geq \lim_{k \rightarrow \infty} \frac{\lambda_{\min}[P_k^{-1}] \theta_k^T \theta_k}{n_I \lambda_{\max}[P_k^{-1}]} \geq 0 \quad (71)$$

$$\Leftrightarrow \lim_{k \rightarrow \infty} \hat{\theta}_k = 0. \quad (72)$$

End of proof.

Remarks: Hereafter, let us make some remarks.

1) Implementation of the proposed algorithm in the case of Wiener-Hammerstein model with time delay. A particular case of the realization (1)–(9) is that

$$b_{i0} = \dots = b_{i, \tau_i - 1} = 0 \quad \text{and} \quad b_{i, \tau_i} \neq 0$$

$$c_{i0} = \dots = c_{i, \tau_i - 1} = 0 \quad \text{and} \quad c_{i, \tau_i} \neq 0 \quad \text{for } i = 1, \dots, r$$

this configuration depends upon the structure selection from the input output data, where τ_i and u_i are times delay related to the i th input.

Polynomials $B_i(q^{-1})$ and $C_i(q^{-1})$ become

$$B_i(q^{-1}) = q^{-\tau_i} (b_{i, \tau_i} + b_{i, \tau_i+1} q^{-1} + \dots + b_{i, \tau_i+1} q^{-1})$$

$$C_i(q^{-1}) = q^{-\tau_i} (c_{i, \tau_i} + c_{i, \tau_i+1} q^{-1} + \dots + c_{i, \tau_i+1} q^{-1})$$

and in the proposed algorithm we replace simply Z_{ik} by $Z_{ik-\tau_i}$ and u_{ik} by $u_{ik-\tau_i}$.

2) Initialization of the algorithm, the mean of the sufficient condition (54) for convergence, and how (52)–(54) can be ensured in the practical case. Indeed, most identification methods developed in the literature are based on a linearized model around a trajectory which is generally given by the last estimation. This approach is available only if the last estimate belongs to a neighborhood of the actual parameter vector, the condition (54) represents then the validity of the linearized model. In practice, if the proposed algorithm is adequately initialized (54) is generally fulfilled. Thus, a bad initialization leads generally to various problems such as convergence to a local minimum, to a wrong estimate or instability [16], [17]. Anyway, there is no approach which can be recommended as being universal to give an analytic solution to this problem. This fact is related to the many different unexpected features that the non linear relation can introduce.

But, for the particular case of the Wiener-Hammerstein model, one way to overcome the initialization difficulty is to write the output signal as a linear parameter representation [5], [6] [11]:

$$y_k = \bar{\phi}_k^T \bar{\theta}_k \quad (73)$$

where $\bar{\phi}_k$ is the data vector which contains only input and output signals and, where $\bar{\theta}_k$ is an over parameter vector. Note that (73) is an equivalent representation of the model given by (33) with an overparameterization. Next to that the use of a least square scheme provides a consistent-like initial parameter vector which may initialize adequately the proposed algorithm, see also [3] and [19].

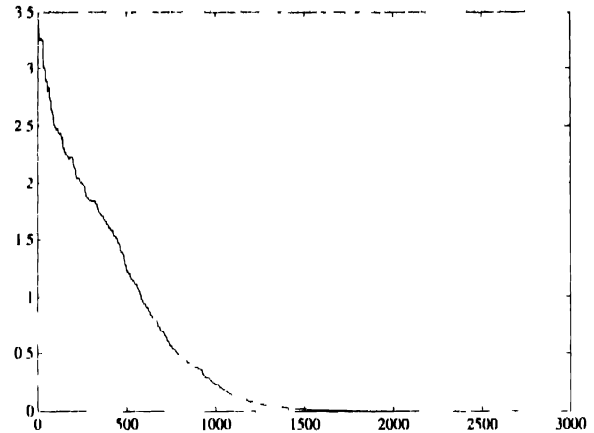


Fig. 2 Rate of convergence of parameter error norm. $\|\theta_k - \theta\|$

Furthermore, (52), (53) represent the persistently exciting conditions, some basic definitions are given in [17] and [19]. However, in practice we must take into account the constraints that physical limitations and safety may impose. Many strategies were developed to achieve a given input spectrum where the amplitude and variance are constrained to a certain interval. Some experiment results are given in [7] and [10].

VII. NUMERICAL EXAMPLE

The numerical example considered here is composed of two subsystems with the nonlinear static polynomials

$$V_{1t} = g_{11} Z_{1t} + g_{12} Z_{1t}^2$$

$$V_{2k} = g_{21} Z_{2k} + g_{22} Z_{2k}^2 + g_{23} Z_{2k}^3$$

and the pulse transfer functions:

$$\frac{B_1}{A_1} = \frac{1 + b_{11} q^{-1}}{1 + a_{11} q^{-1} + a_{12} q^{-2}}$$

$$\frac{C_1}{D_1} = \frac{1 + c_{11} q^{-1} + c_{12} q^{-2}}{1 + d_{11} q^{-1} + d_{12} q^{-2}}$$

$$\frac{B_2}{A_2} = \frac{1}{1 + a_{21} q^{-1}}; \quad \frac{C_2}{D_2} = \frac{1 + c_{21} q^{-1}}{1 + d_{21} q^{-1} + d_{22} q^{-2}}$$

with $a_{11} = 0.5, a_{12} = 0.35, a_{21} = 0.7, b_{11} = 0.1, c_{11} = -0.5, c_{12} = 0.6, c_{21} = 0.8, d_{11} = 0.65, d_{12} = 0.35, d_{21} = 0.45, d_{22} = 0.55, g_{11} = 1.2, g_{12} = -1.8, g_{21} = 0.95, g_{22} = 1.3, g_{23} = -1.1$.

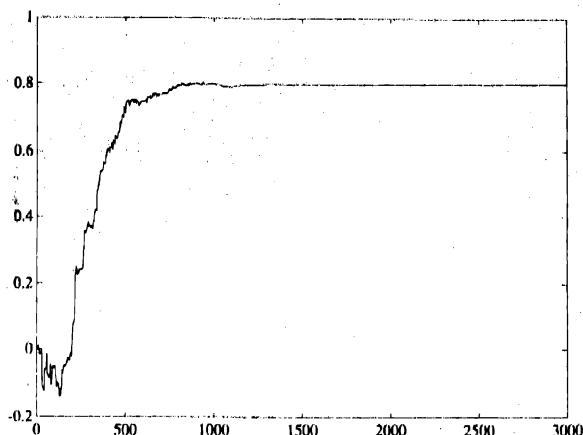
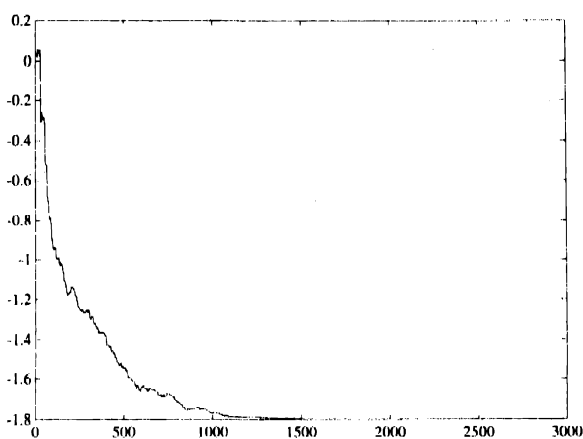
The input signals (u_{1k}) and (u_{2k}) are zero mean white noise sequences with standard deviations $\sigma_1 = 2$ and $\sigma_2 = 2.5$, respectively. In order to show efficiency of the proposed algorithm, the sequence (λ_k) is chosen large enough to fulfill condition (54) and then to enhance convergence of the algorithm, we take

$$\lambda_{k+1} = 10 \hat{f}_k^T P_k \hat{f}_k + 1$$

where the initial estimates θ_0 and P_0 are taken as

$$\hat{\theta}_0 = 0; \quad P_0 = 10^4 I_{16}.$$

Numerical results are given in Figs. 2–4 where the true values of all parameters are reached about 1800 samples. Fig. 2 shows the rate of convergence of parameter error norm: $\|\theta_k - \theta\|$. Owing to a lack of space here, only the estimates of two parameters are given, for example: c_{21} and g_{12} which are represented by Figs. 3 and 4, respectively.

Fig. 3. Parameter estimation of c_{21} .Fig. 4. Parameter estimation of g_{12} .

VIII. CONCLUSION

A simple method for recursive identification of MISO Wiener-Hammerstein model was performed. It is shown, by means of an attractive transformation, that parameters to be estimated are those of each subsystem of the initial and unique realization. Sufficient conditions are given to assure the convergence of the algorithm where the sequence $\{\lambda_k\}$ is introduced to enlarge the interval $[1 - \sqrt{\Delta_{k+1}}, 1 + \sqrt{\Delta_{k+1}}]$ and then to enhance the convergence of the proposed procedure, this was illustrated by means of a numerical example. Finally, some important remarks to deal with the initialization problem, in order to ensure condition (54), are given.

REFERENCES

- [1] S. A. Billings and S. Y. Fakhouri, "Identification of a class of non linear systems using correlation analysis," *IEE*, vol. 125, pp. 691-697, 1978.
- [2] —, "Identification of systems containing linear dynamic and static nonlinear element," *Automatica*, vol. 18, no. 1, pp. 15-26, 1982.
- [3] S. A. Billings and W. S. F. Voon, "A prediction-error and stepwise-regression estimation algorithm for non linear systems," *Int. J. Contr.*, vol. 44, no. 3, pp. 803-822, 1986.
- [4] M. Boutayeb, M. Darouach, H. Rafaralahy, and G. Krzakala, "A new technique for identification of MISO Hammerstein model," in *Proc. Amer. Control Conf. California*, 1993, pp. 1991-1992.
- [5] M. Boutayeb and M. Darouach, "Recursive identification method for Hammerstein model. Extension to the non linear MISO systems," *Contr. Theory Advanced Technol.*, to be published.
- [6] F. H. Chang and R. Luus, "A non-iterative method for identifying nonlinear systems using the Hammerstein model," *IEEE Trans. Automat. Contr.*, vol. 16, pp. 464-468, 1971.
- [7] E. Eskinat, S. H. Johnson, and W. L. Luyben, "Identification of nonlinear models in identification of non linear systems," *Automatica*, vol. 27, no. 2, pp. 255-268, 1991.
- [8] A. Gelb, *Applied Optimal Estimation*. Cambridge, MA: MIT Press, 1974.
- [9] R. Haber and H. Unbehauen, "Structure identification of nonlinear dynamic systems—A survey on input/output approaches," *Automatica*, vol. 26, no. 4, pp. 651-677, 1990.
- [10] M. J. Korenberg, "Identification of biological cascades of linear and static non linear systems," in *Proc. 16th Midwest Symp. Circuit Theory*, 1973.
- [11] M. Kortmann and H. Unbehauen, "Identification methods for nonlinear MISO systems," in *Proc. IFAC 10th Triennial World Congress*, Munich, Germany, 1987, pp. 233-238.
- [12] I. J. Leontaritis and S. A. Billings, "Input-output parametric models for non linear systems. Part I: Deterministic non linear systems," *Int. J. Contr.*, vol. 41, no. 2, pp. 303-328, 1985.
- [13] —, "Input-output parametric models for non linear systems. Part II: Stochastic non linear systems," *Int. J. Contr.*, vol. 41, no. 2, pp. 329-344, 1985.
- [14] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Cambridge, MA: MIT Press, 1983.
- [15] K. S. Narendra and P. G. Gallman, "An iterative method for the identification of nonlinear systems using a Hammerstein model," *IEEE Trans. Automat. Contr.*, vol. AC-11, pp. 546-560, 1966.
- [16] P. Stoica, "On the convergence of an iterative algorithm used for Hammerstein system identification," *IEEE Trans. Automat. Contr.*, vol. AC-26, no. 4, pp. 967-969, 1981.
- [17] P. Stoica and T. Söderström, "Instrumental variable methods for identification of Hammerstein systems," *Int. J. Contr.*, vol. 35, pp. 459-476, 1982.
- [18] K. Yoshine and N. Ishii, "Non-linear analysis of a linear-non-linear system," *Int. J. Syst. Sci.*, vol. 23, no. 4, pp. 623-630, 1992.
- [19] M. Zhao and Y. Lu, "Parameter identification and convergence analysis based on the least-squares method for a class of non linear systems," *Int. J. Syst. Sci.*, vol. 22, no. 1, pp. 33-48, 1991.

Production Rate Control for Failure-Prone Production Systems With No Backlog Permitted

Jian-Qiang Hu

Abstract—Previously, the problem of optimal production rate control for failure-prone production systems has been studied exclusively under the assumption that backlog is permitted. It is well known that when backlog is permitted, the optimal control is usually the hedging point policy. In this note, we consider systems in which backlog is not allowed. We show that the hedging point policy is still optimal. For systems with backlog, it is usually quite straightforward to show that their optimal cost-to-go functions are convex—a key property that is needed for the hedging point policy to be optimal. With no backlog permitted, it becomes much more difficult to establish the convexity property, and the explicit formulas for the optimal hedging point and the optimal cost-to-go functions have to be obtained, based on which the convexity property can then be verified. The method we use in this note to derive these explicit formulas is mainly based on an interesting relationship between the inventory process of the system under the hedging point policy and some stochastic process which is well studied in queueing theory.

Manuscript received January 12, 1994; revised June 1, 1994. This work was supported in part by the National Science Foundation under Grants EID-921211 and DDM-9212368.

The author is with the Manufacturing Engineering Department, Boston University, Boston, MA 02215 USA.

IEEE Log Number 9406988.

I. INTRODUCTION

In this note, we consider the problem of optimal production rate control for failure-prone production systems. This problem has been studied extensively by many researchers in the literature (for a review and a comprehensive list of references on the subject, see a recent monograph by Gershwin [4]). A key assumption used in all the previous work is that backlog of demands is allowed. This assumption usually implies that the instantaneous cost function on inventory level (negative inventory being backlog) is convex (in most cases, it is either piecewise linear or quadratic function of inventory level), which can immediately lead to convexity of optimal cost-to-go functions based on a simple sample path argument. Using the convex property and the dynamic Bellman equation for the optimal cost-to-go functions, one can then easily show that the optimal control is a hedging point policy, in which a (nonnegative) production inventory level of part types is maintained during times of excess capacity availability to hedge against future capacity shortages brought about by machine failures. Intuitively, the hedging point is safety production inventory and is carried to protect the production process from the uncertainty in future capacity availability.

Although the backlog assumption is reasonable for many systems, it may not hold for a large class of production systems in which unsatisfied demands are lost. One such example is a production line in which machines are linked in tandem. In such a system, the demand rate for each machine (except the last one) is simply equal to the production rate of its downstream machine, and obviously, in this case, production inventory level between any two machines can never become negative (i.e., backlog is not permitted). This motivates us to consider the problem of optimal production rate control for systems in which backlog is not permitted. As we shall see, for a system with no backlog permitted, it becomes much more difficult to establish convexity for its optimal cost-to-go functions. This is simply because, in this case, the instantaneous cost function on inventory level is no longer convex, hence the simple sample path argument used before when backlog is permitted cannot be applied. Instead, the explicit formulas for the optimal cost-to-go functions have to be derived.

To derive the optimal cost-to-go functions, we first focus on hedging point policies and try to obtain the optimal hedging point policy and its associated cost-to-go functions. Clearly, if the optimal control policy is a hedging point policy, then the optimal cost-to-go functions are simply equal to the cost-to-go functions associated with the optimal hedging point policy. To obtain the optimal hedging point policy, we use a method which has recently been developed by Hu and Dong [5], [6] to study the systems with backlog permitted. The basic idea of the method is to establish a relationship between the inventory processes of failure-prone production systems and some well-studied stochastic processes in queueing systems. In our case, we show that the inventory process of the system under a hedging point policy is related to the workload process of a single-node queueing system with limited workload. Based on this relationship and existing results in queueing theory, we can obtain the probability distribution function of the inventory process, from which we can then find the optimal hedging point. Finally, with the optimal hedging point policy and its cost-to-go functions, we can use the verification theorem to verify that the optimal hedging point policy is indeed optimal overall. We point out that it is also possible for one to obtain the probability distribution function of the inventory process for our system based on a different approach used in [1], [2], [8], in which a system of two differential equations needs to be solved under two boundary conditions. However, the method used in [1], [2], [8] can only be applied to Markov systems (i.e., both machine up and down times have to be exponentially distributed), while our method is applicable to non-Markov systems as well.

The remainder of the note is organized as follows. In Section II, the problem of optimal production rate control for the failure-prone production system with no backlog permitted is formulated as an optimal control problem whose optimality condition is given by a dynamic Bellman equation. In Section III, we establish an important relationship between the inventory process of the system under a hedging point policy and the workload process of a single-node queue with bounded workload. Based on this relationship, we obtain the probability distribution function for the inventory process. We then proceed in Section IV to derive the explicit formulas for the optimal hedging point, the optimal value of the cost function, and the optimal cost-to-go functions. In Section V, we present the verification theorem, based on which we prove that the optimal hedging point policy obtained in Section IV is the optimal control policy. Finally, some discussions and future research directions are presented in Section VI.

II. PROBLEM FORMULATION

We consider a production system which has a single machine and produces a single part type. The system tries to meet a constant demand rate d , and backlog is not permitted, i.e., unsatisfied demands are lost. The machine has two states: up and down. When the machine is up, it can produce at any rate between zero and a maximum rate r . We assume that $r > d$. The machine state changes in continuous time according to a homogeneous Markov process: the state changes from down to up at a rate q_0 and from up to down at a rate q_1 (i.e., both machine up and down times are exponentially distributed with rates q_1 and q_0 , respectively). We use $i \in \{1, 0\}$ to denote the state of the machine, where one corresponds to up and zero to down. Denote the production inventory at time t by $x(t) (\geq 0)$. Let $i(t)$ be the state of the machine at time t , and let $u(t)$ be the controlled production rate of the machine at time t . Then $x(t)$ is characterized by the following differential equation:

$$\frac{dx(t)}{dt} = \begin{cases} u(t) - d, & \text{if } i(t) = 1 \text{ or } x(t) > 0 \\ 0, & \text{if } i(t) = 0 \text{ and } x(t) = 0 \end{cases} \quad (1)$$

where $u(t) \in [0, r]$ when $i(t) = 1$ ($u(t) \in [d, r]$ if $x(t) = 0$), and $u(t) = 0$ when $i(t) = 0$. The objective is to find a stationary, feedback control law, $u(x, i)$, so as to minimize the following long-run expected average cost:

$$J = \lim_{T \rightarrow \infty} E \left\{ (1/T) \int_0^T (cx(t) + c_0 1(x(t) = 0, i(t) = 0)) dt \right\} \quad (2)$$

where c is the unit holding cost for inventory, c_0 is the unit shortage cost for unsatisfied demand, and $1(\cdot)$ is the indicator function. It should be pointed out that the instantaneous cost in the objective function, $g(x, i) = cx + 1(x = 0, i = 0)$, is not a convex function with respect to x .

Under suitable regularity conditions imposed on the control (e.g., [7]), the optimal feedback control of the system is stationary and characterized by the following Bellman equations:

$$\min_{u: 0 \leq u \leq r} \left\{ \frac{dV_1(x)}{dx} (u - d) \right\} - q_1 [V_1(x) - V_0(x)] + cx - J^* = 0, \quad \text{for } x \geq 0 \quad (3)$$

$$-\frac{dV_0(x)}{dx} d - q_0 [V_0(x) - V_1(r)] + cx - J^* = 0, \quad \text{for } x > 0 \quad (4)$$

where $V_i(x)$ ($i = 1, 0$) are (differential) cost-to-go functions and J^* is the minimum long-run expected average cost associated (i.e.,

$J^* = \min J$). According to (3), if $V_1(x)$ is convex with minimum point z^* , then the optimal production rate satisfies

$$u^*(x, i) = \begin{cases} 0, & \text{if } x > z^* \text{ and } i = 1 \\ d, & \text{if } x = z^* \text{ and } i = 1 \\ r, & \text{if } x < z^* \text{ and } i = 1 \end{cases} \quad (5)$$

which is hedging point policy, and z^* is the hedging point. If the instantaneous cost function were convex, then one could easily show that $V_1(x)$ is also convex based on a simple sample argument ([10]). However, as already mentioned above, the instantaneous cost function in our case is not convex, hence the explicit formula for $V_1(x)$ has to be obtained in order to establish its convexity.

III. HEDGING POINT POLICY

In this section, we focus exclusively on the hedging point policy defined by (5) with hedging point z . We want to derive the steady-state probability distribution function for the inventory process under the hedging point policy. Our method consists of two basic steps. First, we establish a relationship between the inventory process and the workload process of a single-node queueing system with bounded workload. Second, we use this relationship to derive the probability distribution function for the inventory process based on existing results in queueing theory and a level crossing technique. This method was first used by Hu and Dong [5], [6] to study systems with backlog. It is worth pointing out that a different approach proposed in [1], [2], [8] can also be used to obtain the probability distribution function in our case. The basic idea of their method is to solve a set of differential equations. However, we note that the method of [1], [2], [8] can only be applied to Markov systems, while ours is applicable to non-Markov systems as well (i.e., the machine up and down times do not have to be exponentially distributed).

To start, we take a close look at a sample path of the system under the hedging point policy. For simplicity, let us assume $x(0) = z$ and $i(0) = 0$. Denote by $t_{d,n}$ the length of the n th down time and by $t_{u,n}$ the length of the n th up time. Therefore, $t_{d,n}$'s and $t_{u,n}$'s are i.i.d. exponential random variables with rates q_0 and q_1 . Let $T_{d,n}$ be the beginning of the epoch of the n th time the machine is down and $T_{u,n}$ be the beginning of the epoch of the n th time the machine is up, i.e.,

$$T_{d,n} = \sum_{i=1}^{n-1} (t_{d,i} + t_{u,i}) \quad \text{and} \quad T_{u,n} = T_{d,n} + t_{d,n}. \quad (6)$$

To simplify notation, we denote

$$x_{d,n} \triangleq x(T_{d,n}) \quad \text{and} \quad x_{u,n} \triangleq x(T_{u,n}).$$

We then have the following recursive equations for $x_{d,n}$ and $x_{u,n}$:

$$x_{d,n} = \max(x_{d,n-1} - t_{d,n}d, 0) \quad (7)$$

$$x_{u,n+1} = \min(z, x_{u,n} + t_{u,n}(r-d)) \quad (8)$$

with $x_{d,1} = z$. Equation (7) represents the unique dynamics when the machine is down, where "max" means that the inventory level cannot fall below zero (i.e., backlog is not permitted). The "min" in (8) represents the hedging point policy, namely, when the machine is up, the inventory level increases at rate $r-d$ until either it hits the level z where it remains until the machine breaks down, or the machine breaks down before it hits the level z . The trajectory of $x(t)$ is illustrated in Fig. 1.

Define the following two random time transformations:

$$\omega_d(t) = \int_0^t 1(i(s) = 0) ds$$

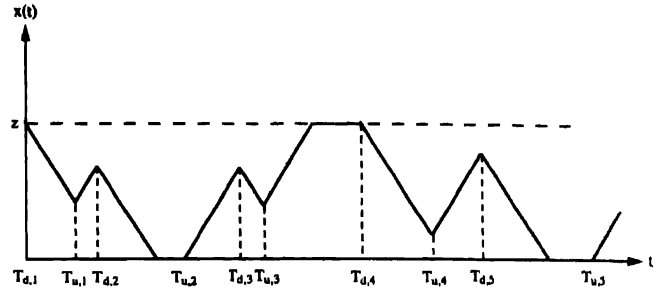


Fig. 1. The trajectory of $x(t)$.

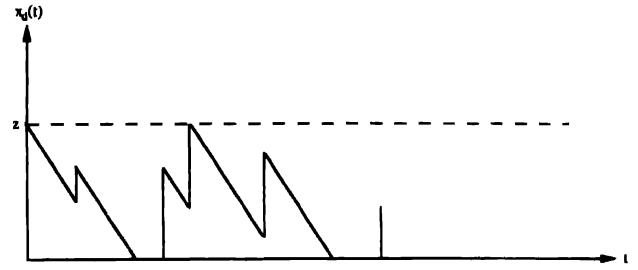


Fig. 2. The trajectory of $x_d(t)$.

$$\phi_d(t) = \int_0^t 1(i(s) = 1) ds$$

and their inverse functions

$$\tau_d(t) = \phi_d^{-1}(t) = \inf \{ \tau : \phi_d(\tau) > t \}$$

$$\tau_u(t) = \phi_u^{-1}(t) = \inf \{ \tau : \phi_u(\tau) > t \}.$$

Define also two processes $\{x_d(t); t \geq 0\}$ and $\{x_u(t); t \geq 0\}$ as

$$x_d(t) = x(\tau_d(t)) \quad \text{and} \quad x_u(t) = x(\tau_u(t)).$$

Intuitively, $\phi_d(t)$ [respectively, $\phi_u(t)$] is the total down (respectively, up) time of the machine during the time interval $[0, t]$. The process $\{x_d(t); t \geq 0\}$ (respectively, $\{x_u(t); t \geq 0\}$) corresponds to the part of the process $\{x(t); t \geq 0\}$ when the machine is down (respectively, up). The process $\{x_d(t); t \geq 0\}$ is illustrated in Fig. 2.

It is not too difficult to verify that $x_d(t)$ is in fact the same as the workload process of an $M/M/1$ queue with bounded workload, in which the interarrival times are $\{t_{d,n}; n = 1, 2, \dots\}$, the service requirements are $\{t_{u,n}(r-d); n = 1, 2, \dots\}$, the service rate is d , and the upper bound on the workload process is z . Hence, the steady-state probability distribution function of the process $\{x_d(t); t \geq 0\}$ is given by

$$F_{x_d}(x) = \begin{cases} 1, & x \geq z \\ \frac{1-\rho - \mu(1-\rho)^x}{1-\rho - \mu(1-\rho)^z}, & 0 \leq x < z \\ 0, & x < 0 \end{cases} \quad (9)$$

where

$$\mu = \frac{q_1}{r-d} \quad \text{and} \quad \rho = \frac{q_0(r-d)}{q_1 d}.$$

(See Takács [9].) Next we use $F_{x_d}(x)$ to calculate $F_{x_d}(x)$, the steady-state probability distribution function of the process $\{x_d(t); t \geq 0\}$. We use $f_{x_d}(x)$ and $f_{x_u}(x)$ to denote the density functions of $F_{x_d}(x)$ and $F_{x_u}(x)$, respectively.

Define $D_x(t)$ to be the downcrossing counting process of $x(t)$ at level x during the interval $[0, t]$, and $U_x(t)$ the upcrossing counting process. Define $t_0 = 0$, and

$$t_n = \inf \{ t : x(t) = z, i(t) = 0, t > t_{n-1} \} \quad \text{for } n \geq 1.$$

By definition, it is clear that $\{x(t), t \geq 0\}$ is a regenerative process with regenerative points $\{t_n, n \geq 0\}$. Based on the regenerative theorem, we have

$$F_{\lambda}(x) = \frac{E\left[\int_0^{t_1} 1(x(s) < x) ds\right]}{E[t_1]}.$$

Note that

$$\int_0^{t_1} 1(x(s) < x + \delta x) ds - \int_0^{t_1} 1(x(s) < x) ds$$

is the total time that the process $x(t)$ takes value between the interval $[x, x + \delta x)$ during $[0, t_1]$, which is also equal to

$$\frac{U_r(t_1)\delta x}{r-d} + \frac{D_r(t_1)\delta x}{d} \quad \text{for } 0 < x < z.$$

We further note that $U_r(t_1) = D_r(t_1)$ for $0 < x < z$. Therefore, we have

$$\frac{d}{dx} E\left[\int_0^{t_1} 1(x(s) < x) ds\right] = \frac{r}{(r-d)d} E[D_r(t_1)].$$

This gives us

$$\begin{aligned} f_{\lambda}(x) &= \frac{dF_{\lambda}(x)}{dx} = \frac{r}{(r-d)d} \frac{E[D_r(t_1)]}{E[t_1]} \\ &= \frac{r}{r-d} \frac{E[\phi_d(t_1)]}{E[t_1]} \frac{E[D_r(t_1)]}{dE[\phi_d(t_1)]} \\ &= \frac{r}{r-d} \frac{q_1}{q_0 + q_1} \frac{E[D_r(t_1)]}{dE[\phi_d(t_1)]} \quad \text{for } 0 < x < z. \end{aligned} \quad (10)$$

The above method, which we used to derive (10), is the standard analysis of level crossings (e.g., see Cohen [3, p. 363-373]). Using the same method, we can also show that the most right-hand side of (10) is equal to $f_{\lambda,d}(x)$. Therefore, we have

$$f_{\lambda}(x) = \frac{r}{r-d} \frac{q_1}{q_0 + q_1} \frac{\rho\mu(1-\rho)e^{-\mu(1-\rho)z}}{1-\rho} \quad \text{for } 0 < x < z. \quad (11)$$

We now calculate $F_{\lambda}(0)$ and $F_{\lambda}(z)$. Again, using the regenerative theorem, we have

$$\begin{aligned} F_{\lambda}(0) &= \frac{E\left[\int_0^{t_1} 1(x(s) = 0) ds\right]}{E[t_1]} \\ &= \frac{E[\phi_d(t_1)]}{E[t_1]} \frac{E\left[\int_0^{\phi_d(t_1)} 1(x_d(\phi_d(s)) = 0) d\phi_d(s)\right]}{E[\phi_d(t_1)]} \\ &= \frac{q_1}{q_0 + q_1} F_{\lambda,d}(0) = \frac{q_1}{q_0 + q_1} \frac{1-\rho}{1-\rho e^{-\mu(1-\rho)z}} \end{aligned}$$

and

$$F_{\lambda}(z) = 1 - F_{\lambda}(0) - \int_0^z f_{\lambda}(x) dx = \frac{q_0}{q_0 + q_1} \frac{(1-\rho)e^{-\mu(1-\rho)z}}{1-\rho e^{-\mu(1-\rho)z}}.$$

Finally, we calculate the long-run expected average cost J , associated with the hedging point policy. Based on (2), it is quite clear that

$$\begin{aligned} J &= c \int_0^z x f_{\lambda}(x) dx + c_0 F_{\lambda}(0) \\ &= \frac{1}{(q_0 + q_1)(1-\rho e^{-\mu(1-\rho)z})} \left[c_0 q_1 (1-\rho) \right. \\ &\quad \left. + c\rho \left[\frac{r(1-e^{-\mu(1-\rho)z})}{1-\rho} - (q_0 + q_1)ze^{-\mu(1-\rho)z} \right] \right]. \end{aligned} \quad (12)$$

IV. OPTIMAL COST-TO-GO FUNCTIONS

To obtain the optimal cost-to-go functions, let us assume throughout this section that the optimal control is the hedging point policy. This will be verified later in the next section. Let z^* be the optimal hedging point, then (12) tells us that it must satisfy the following equation:

$$h(z) \triangleq \frac{c(q_0 + q_1)z}{1-\rho} + \frac{c}{(1-\rho)^2} [d - (r+d)\rho + \rho(r-d+\rho d)e^{-\mu(1-\rho)z}] - c_0 q_1 = 0 \quad (13)$$

(this is simply due to $dJ/dz = 0$), and furthermore,

$$Jz^* = cz^* + \frac{cd}{q_1 + q_1}. \quad (14)$$

It is worth noting that z^* is the unique solution of $h(z) = 0$ (because $dh(z)/dz > 0$). However, $z^* = 0$ if the unique solution of $h(z) = 0$ is negative (which is the case if $h(0) = cd - c_0 q_1 > 0$). Since we have assumed that the optimal control is the hedging point policy, we have $J^* = J_{z^*}$, and furthermore, the Bellman equations (3) and (4) reduce to

$$\frac{dV(x)}{dx} = \begin{cases} A_1 V(x) - b_1(J^* - cx), & \text{for } 0 < x \leq z^* \\ A_2 V(x) - b_2(J^* - cx), & \text{for } x \geq z^* \end{cases} \quad (15)$$

where

$$A_1 = \begin{pmatrix} \frac{q_1}{r-d} & -\frac{q_1}{r-d} \\ \frac{q_0}{d} & -\frac{q_0}{d} \end{pmatrix}, \quad A_2 = \begin{pmatrix} -\frac{q_1}{d} & \frac{q_1}{d} \\ \frac{q_0}{d} & -\frac{q_0}{d} \end{pmatrix}.$$

$$V(x) = \begin{pmatrix} V_1(x) \\ V_0(x) \end{pmatrix}, \quad b_1 = \begin{pmatrix} -\frac{1}{r-d} \\ \frac{1}{d} \end{pmatrix}, \quad b_2 = \begin{pmatrix} \frac{1}{d} \\ \frac{1}{d} \end{pmatrix}$$

with boundary conditions

$$c_0 - J^* + q_0(V_1(0) - V_0(0)) = 0 \quad (16)$$

$$cz^* - J^* - q_1(V_1(z^*) - V_0(z^*)) = 0. \quad (17)$$

Solving (15), we have, for $0 \leq x \leq z^*$,

$$\begin{aligned} V(x) &= \left(I - \frac{A_1}{k_1} + \frac{A_1}{k_1} e^{k_1 x} \right) V(0) + \left(\frac{A_1}{k_1} - I \right) b_1 \left(J^* x - \frac{1}{2} cx^2 \right) \\ &\quad + \frac{A_1 b_1}{k_1^2} \left[J^* (1 - e^{k_1 x}) - cx + \frac{c}{k_1} (e^{k_1 x} - 1) \right] \end{aligned} \quad (18)$$

where $k_1 = \mu(1-\rho)$, and for $x \geq z^*$,

$$\begin{aligned} V(x) &= \left(I - \frac{A_2}{k_2} + \frac{A_2}{k_2} e^{k_2(x-z^*)} \right) V(z^*) \\ &\quad - \left(J^*(x-z^*) - \frac{1}{2} c(x-z^*)^2 \right) b_2 \end{aligned} \quad (19)$$

where $k_2 = -(q_0 + q_1)/d$.

Based on (18) and also using (16), we have for $0 \leq x \leq z^*$

$$\begin{aligned} V_1(x) &= V_1(0) + \frac{q_1}{k_1(r-d)q_0} (J^* - c_0)(e^{k_1 x} - 1) \\ &\quad - \frac{q_0 + q_1}{k_1 d(r-d)} \left(J^* x - \frac{1}{2} cx^2 \right) \\ &\quad - \frac{q_1 r}{k_1^2 d(r-d)^2} \left[J^* (1 - e^{k_1 x}) - cx + \frac{c}{k_1} (e^{k_1 x} - 1) \right] \end{aligned} \quad (20)$$

and

$$\frac{dV_1(i)}{di} = \frac{q_1(q_0 + q_1)}{k_1 q_0 (i - d)^2} J^* e^{k_1 i} - \frac{c_0 q_1}{(i - d) q_0} e^{k_1 i} - \frac{q_0 + q_1}{k_1 d (i - d)} (J^* - c_1) + \frac{c_1 q_1}{k_1 d (i - d)^2} (1 - e^{k_1 i}) \quad (21)$$

Substituting c_0 and J^* from (13) and (14) into (21), we obtain

$$\frac{dV_1(i)}{di} = \frac{c(q_0 + q_1)(i - d)}{q_1^2 d (1 - \rho)^2} (1 - e^{-k_1(i - i^*)}) - \frac{c(q_0 + q_1)}{k_1 d (i - d)} (J^* - c_1) \quad \text{for } 0 \leq i \leq i^* \quad (22)$$

Similarly, by combining (19), (17), and (14), we obtain for $i \geq i^*$

$$V_1(i) = V_1(i^*) + \frac{J^* - c_1}{k_1 d} (e^{k_1(i - i^*)} - 1) + \frac{1}{d} \left[\frac{1}{2} (e^{k_1(i - i^*)} - e^{k_1(i^* - i^*)}) - J^* (i - i^*) \right] \quad (23)$$

and

$$\frac{dV_1(i)}{di} = -\frac{c}{q_0 + q_1} (e^{k_1(i - i^*)} - 1) + \frac{c}{d} (i - i^*) \quad (24)$$

V THE VERIFICATION THEOREM

In this section we verify that the optimal hedging point policy we obtained in the previous section is the optimal control. We first need the following verification theorem.

Verification Theorem For the system under consideration suppose the control $u^*(i, i)$ is defined by (5) where i^* is given by (13). If there exist continuously differentiable functions $V(i)$ on $i \in (0, \infty)$ and a constant J^* such that the Bellman equations (3) and (4) are satisfied and $V(i) < ai^* + b$ (where a and b are two constants) then $u^*(i, i)$ is the optimal control which minimizes the long run expected average cost defined by (2).

The proof of the verification theorem is essentially the same as the one given in [2] which uses the Dynkin formula. Hence we will not repeat it here. For the optimal cost-to-go function $V_1(i)$ which we obtained in the previous section we can easily verify [based on (22) and (24)] that it is a convex function with minimum J^* and also it is continuously differentiable. Hence based on the above verification theorem the optimal hedging point policy is indeed optimal.

VI CONCLUSION

We proved that even when backlog is not permitted, the optimal control for the failure-prone production system remains to be a hedging point policy. We demonstrated that the traditional sample path argument is no longer valid in proving that the optimal cost-to-go functions are convex functions. Instead the explicit formulas have to be obtained. To obtain the optimal hedging point policy and the optimal cost to go functions, we first established a relationship between the inventory process of the system under a hedging point policy and the workload process of a single-node queueing system with limited workload, and then used it to obtain the probability distribution function of the inventory process based on some existing results in queueing theory and the level crossing technique.

An immediate application of our results in this note is to failure-prone production systems with many machines connected in network. One way to study the optimal control problem for these complex

systems is to decompose them into many simple subsystems systems with only one machine. The interaction among these simple subsystems can be formulated through the total demand rate to each subsystem. Clearly, for some subsystems backlog of demands will not be permitted, especially if they are suppliers to other subsystems. One example is a production line in which machines are linked together in tandem. In such a system the demand rate for each machine (except the last one) is simply equal to the production rate of its downstream machine hence backlog will not be permitted. Our results can give optimal instantaneous production rates for those subsystems in which backlog is not permitted. Hence, they can be used to develop optimal or near-optimal designs on production rate control for complex systems when combined with other methods.

ACKNOWLEDGMENT

The author wishes to thank the two anonymous referees for their comments and suggestions.

REFERENCES

- [1] R. Akella and P. R. Kumar, "Optimal control of production rate in a failure-prone manufacturing system," *IEEE Trans. Automat. Contr.*, vol. 31, pp. 116-126, 1986.
- [2] J. Bielecki and P. R. Kumar, "Optimality of zero inventory policies for unreliable manufacturing systems," *Operat. Res.*, vol. 36, pp. 532-541, 1988.
- [3] J. M. Cohen, *The Single Server Queue*, rev. ed., New York: North-Holland, 1982.
- [4] S. B. Gershwin, *Manufacturing Systems Engineering*, Englewood Cliffs, NJ: Prentice Hall, 1994.
- [5] J. Q. Hu and D. Xiang, "A queueing equivalence to optimal control of a manufacturing system with failures," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 499-502, 1993.
- [6] ———, "Optimal control for systems with deterministic production cycles," *IEEE Trans. Automat. Contr.*, to appear, 1995.
- [7] R. Rishel, "Dynamic programming and minimum principles for systems with jump Markov distributions," *SIAM J. Contr.*, vol. 36, no. 2, Feb. 1975.
- [8] A. Sharifnia, "Optimal production control of a manufacturing system with machine failures," *IEEE Trans. Automat. Contr.*, vol. AC-33, pp. 620-625, July 1988.
- [9] I. Takács, "A single server queue with limited virtual waiting time," *J. Appl. Probability*, vol. 11, pp. 612-617, 1974.
- [10] J. N. Tsitsiklis, "Convexity and characterization of optimal policies in a dynamic routing problem," *J. Optim. Theory Appl.*, vol. 44, no. 1, Sept. 1984.

Relative Stability of a Linear Time-Varying Process with First-Order Nonlinear Time-Varying Feedback

J. S. Ansari

Abstract—A method of establishing sufficient conditions for the stability of a system consisting of a time-varying linear process and a first-order nonlinear time-varying feedback is presented. Rate of decay of unforced response and BIBO stability can also be determined. The governing equation of the system is represented in the form of an integro-differential equation, so that a modified form of the Lyapunov–Razumikhin-type condition can be applied. The first part of the note deals with stability of vector functional equations. The results obtained are applied to feedback systems. An example is worked out to show the applicability of the results obtained.

1. INTRODUCTION

Methods such as circle criterion and Popov's criterion are available for establishing sufficient condition for the stability of systems consisting of a linear process and zero-order nonlinear feedback [1], [2]. No general results are available for linear time-varying process with first-order nonlinear, time-varying feedback systems. The method presented here is applicable to such systems. In particular, linear processes with an arbitrarily varying parameter can also be considered. As such, the results may be useful for studying stability of adaptive control systems.

The note is in two parts. The first part presents a method of establishing an upper bound on the norm of the solution of a vector functional differential equation. The results obtained are then applied to a linear time-varying process with a first-order nonlinear, time-varying feedback.

The note presents two points. It shows that the Lyapunov–Razumikhin theorem, which is essentially applicable to functional equations, can be applied to feedback systems governed by ordinary differential equations. This is done by rewriting the system equations in integro-differential form. Some systems which cannot be analyzed by other methods can be studied by this method. The other contribution is that the Lyapunov–Razumikhin theorem has been extended to included inputs and, furthermore, a method of establishing the rate of decay of solutions is also included.

II. LYAPUNOV–RAZUMIKHIN-TYPE THEOREM

Let

$$\dot{y}(t) = f(y_t, t) \quad (1)$$

represent a vector functional differential equation of the retarded type [3], where $y(t)$ is an $(n \times 1)$ vector, and y_t in the argument of f implies dependence of f on $y(s)$, $t - T(t) \leq s \leq t$. Function $[t - T(t)]$ is nondecreasing. Interval $[t - T(t), t]$ is represented by $I(t)$. $\|y\|$ represents Euclidian norm in R^n . Equation (1) is such that $y = 0$ is a solution. If $y(s)$ is given in the initial interval, $I(t_0)$, then $y(t)$ for $t > t_0$ is determined by equation (1). It can be shown that the solution is continuous. A scalar function $V(t, y)$ is chosen such that $w(\|y\|) \leq V(t, y) \leq v(\|y\|)$, where w and v are increasing functions, and $w(0)$ and $v(0)$ are zero. If V tends to zero with time, then it follows that $\|y\|$ also tends to zero.

Lyapunov–Razumikhin sufficient conditions for stability require that [4]

$$\dot{V}(t, p(t)) \leq -\alpha < 0$$

for all continuous test functions, p , that satisfy the inequality

$$V(s, p(s)) < V(t, p(t)) + \delta, \quad s \in I(t).$$

A modified form of a Lyapunov–Razumikhin-type condition is presented here in the form of a theorem. The theorem gives sufficient conditions under which $V(t, y(t)) < H(t)$, for all $t > t_0$; where $H(t)$ is a function of our choice. By choosing H as an exponentially decaying function, we can establish the rate of decay of the response, which is not possible through the original condition. Moreover, the Lyapunov–Razumikhin theorem had not been used to study stability of finite-order linear processes with nonlinear feedback, because it is not applicable to ordinary differential equations. In order to apply a Lyapunov–Razumikhin-type condition, we shall use convolution to represent the governing equation in an integro-differential equation form.

Let $y(t)$ be a solution of equation (1), and $V(t, y(t))$ be a scalar function fulfilling the following conditions.

i)

$$w(\|y\|) \leq V(t, y) \leq v(\|y\|) \quad (2)$$

for $\|y\| \leq Y_0$, $t > -T(t_0)$ where $w(s)$ and $v(s)$ are increasing functions of s for $0 \leq s \leq Y_0$, and $w(0) = v(0) = 0$.

ii) $V(t, y)$ is differentiable with respect to t and by all elements y_i of the vector y . We denote

$$\left(\frac{\partial V}{\partial y_1} + \frac{\partial V}{\partial y_2} + \cdots + \frac{\partial V}{\partial y_n} \right) T = \frac{\partial V}{\partial y}.$$

iii) $\partial V / \partial t$ and all $\partial V / \partial y_i$ are continuous with respect to t and y_i , for $t > t_0$ and $\|y\| \leq Y_0$.

Theorem. If for a continuous function $H(t) \leq v(\|y_0\|)$, with continuous derivative,

$$\lim_{s \rightarrow t} \left[\frac{\partial V(s, p)}{\partial s} + \frac{\partial V}{\partial p} \cdot f(p, s) \right] < \dot{H}(t) \quad \forall t > t_0 \quad (3)$$

For all test functions p , belonging to set $P(t)$, where $P(t)$ is the set of all vector continuous functions for which the following conditions are satisfied:

$$V(t, p(t)) = H(t) \quad (4)$$

$$V(s, p(s)) < H(s), \quad \text{for } t - T(t) \leq s < t \quad (5)$$

then

$$V(t, y(t)) < H(t), \quad \text{for all } t > t_0 \quad (6)$$

provided the initial conditions are such that

$$V(s, y(s)) < H(t_0), \quad \text{for } s \in I(t_0). \quad (7)$$

Proof: It is given that initially

$$V(t_0, y(t_0)) < H(t_0). \quad (8)$$

Suppose there exists a t_1 , such that t_1 is the smallest value of $t > t_0$ for which

$$V(t_1, y(t_1)) = H(t_1). \quad (9)$$

It follows that

$$V(t, y(t)) < H(t), \quad t_0 - T(t_0) \leq t < t_1 \quad (10)$$

Manuscript received May 4, 1993; revised October 20, 1993.

The author is with the Department of Mechanical Engineering, Osmania University, Hyderabad 500 007, India.

IEEE Log Number 9406989.

and that $y(t)$ belongs to the set $P(t_1)$, where P is defined by conditions (4) and (5). Hence, we can substitute y in place of p in condition (3) to get

$$\lim_{s \rightarrow t_1} \left[\frac{\partial V(s, y)}{\partial s} + \frac{\partial V}{\partial y} f(y, s) \right] < H(t_1) \quad (11)$$

According to (1), $f(y, s)$ is equal to $q(s)$. Hence the left hand side of inequality (11) is equal to V and inequality (11) can be written as

$$\lim_{s \rightarrow t_1} V(s, y(s)) < H(t_1) \quad (12)$$

At the same time it follows from relations (9) and (10) that

$$\lim_{s \rightarrow t_1} V(s, y(s)) \geq H(t_1) \quad (13)$$

which contradicts inequality (12). Hence assumption (9) is not true and V is bounded by H .

III. APPLICATION TO FEEDBACK SYSTEMS

Consider a system consisting of a time varying linear n th order asymptotically stable plant coupled with nonlinear time varying feedback. Let the plant be governed by

$$\dot{y}(t) = \int_0^t q(t-t')u(t')dt' + \phi(t, Y) \quad (14)$$

where q is the impulse response function and $\phi(t, Y)$ is the initial condition response and Y is the $(n \times 1)$ vector representing the initial state. For convenience we shall write (14) as

$$\dot{y} = q(u, Y) \quad (14a)$$

Let the feedback be given by

$$u + F(t, u) = \lambda(t, (r - y)) \quad (15)$$

where r is the reference input. Functions $F(t, u)$ and $\lambda(t, v)$ are nondecreasing with respect to u and $\lambda(t, 0)$ and $F(t, 0)$ are identically zero.

Combining (14a) and (15) gives

$$\dot{u} = -F(t, u) + \lambda(t, (r - y(u, Y))) \quad (16)$$

Let us choose

$$V(t, u) = u^2 \quad (17)$$

and

$$H(t) = B e^{-\lambda t} \quad (18)$$

Since the linear plant is asymptotically stable we can find a $\Phi(t)$ such that the initial condition response

$$|\phi(t, Y)| < \epsilon \Phi(t) \quad t \geq 0 \quad (19)$$

if

$$\|Y\| < \epsilon \quad (20)$$

Let the input be bounded by

$$|r(t)| < \delta \quad t \geq t_0 \quad (21)$$

Application of the above theorem to (16) shows the following

Corollary 1 If

$$\begin{aligned} \pm B e^{-\lambda t} \lambda \left[t \pm \left(\delta + \epsilon \Phi(t) + B \int_0^t |q(t-t')| e^{-\lambda t'} dt' \right) \right] \\ < -\lambda + \frac{F(t \pm B e^{-\lambda t})}{\pm B e^{-\lambda t}} \end{aligned} \quad (22)$$

then

$$|u(t)| < B e^{-\lambda t} \quad t < T$$

where T can be infinity provided

$$\|Y\| < \epsilon, |r(t)| < \delta \quad \text{and} \quad |u(t)| < B \quad \text{for } t \in I(t_0)$$

Corollary 2 If for any $B > 0$ and $\delta = 0$ we can find positive λ and ϵ to satisfy condition (22) for all $t > t_0$ then the unforced solution $y = u = 0$ is asymptotically stable.

Corollary 3 If for any $B > 0$ and $\lambda = 0$ we can find positive δ and ϵ to satisfy condition (22) for all $t > t_0$ then the system is stable in the sense of bounded input bounded output (BIBO).

Remark Corollaries follow immediately and obviously from the theorem. Therefore proofs are not included.

A Example

Let

$$\dot{y} + (a + b)y + ab y = cu \quad (23)$$

and

$$u + F(t, u) = \lambda(t, (r - y)) \quad (24)$$

where $b > a > 0$, $c > 0$, $\lambda(t, 0) = F(t, 0) = 0$, $F(t, u)$ and $\lambda(t, v)$ are nondecreasing with respect to u . Let initial conditions be such that

$$|u(0)| < B, \quad \|\lambda(0)\| \leq \epsilon$$

where $\lambda(0) = (y(0) - y(0)/b)e^{-t}$. Let $|r(t)| \leq \delta$, $t \geq 0$.

Evaluating q and Φ and substituting in inequality (22) we get

$$\begin{aligned} \pm B e^{-\lambda t} \lambda \left[t \pm \left(\delta + \frac{B e^{-\lambda t} \exp(-\lambda t)}{(a - \lambda)(b - \lambda)} \right) \right. \\ \left. - \left(\frac{B e^{-\lambda t}}{a - \lambda} - \sqrt{2b\epsilon} \right) \frac{\exp(-\lambda t)}{b - a} \right. \\ \left. + \left(\frac{B e^{-\lambda t}}{b - \lambda} - \sqrt{2a\epsilon} \right) \frac{\exp(-\lambda t)}{b - a} \right] < -\lambda + \frac{F(t \pm B e^{-\lambda t})}{\pm B e^{-\lambda t}} \end{aligned} \quad (25)$$

If $a = 2$, $b = 3$, $c = 1$ and

$$\frac{F(t \pm u)}{u} \geq 10.5 \quad t \geq 0$$

and we choose $\lambda = 0.5$, $\delta = 0$, $\epsilon = B/6\sqrt{2}$ then inequality (25) becomes

$$\frac{F(t \pm u)}{\pm u} < \frac{37.5}{1 - 0.625e^{-t} + 2.75e^{-t}} \quad (26)$$

where

$$u = 0.266B e^{-0.5t} - 0.1625B e^{-t} + 0.75B e^{-1.5t}$$

If inequality (26) is satisfied for all $t > 0$ then

$$|u(t)| < B e^{-0.5t} \quad t \geq 0$$

BIBO Stability: If the values of a , b , c , and F are as chosen above, and we choose $\lambda = 0$, $\delta = B/2$, then inequality (25) yields

$$\frac{N(t, \pm x)}{\pm x} < \frac{15}{1 + e^{-3t/4}} \quad (27)$$

where

$$x = \delta + (B/6) + (2Be^{-3t/4}/3).$$

If condition (27) is satisfied, then

$$|u(t)| < B \quad t \geq 0$$

provided

$$|r(t)| \leq B/2 \quad t \geq 0.$$

B. Comparison to the Result Obtained Through the Original Lyapunov–Razumikhin Theorem

It is of interest to note that the application of the original Lyapunov–Razumikhin condition to (23) and (24) gives the following sufficient condition for asymptotic stability of the unforced solution:

$$\pm B^{-1} N \left[\frac{\pm c B}{b-a} \left(\frac{b-a}{ab} - \frac{e^{-at}}{a} - \frac{e^{-bt}}{b} \right) \right] < \frac{F(\pm B, t)}{\pm B} - \alpha \quad \times \forall B, B_0 \geq B > 0 \quad (28)$$

where α is positive and as small as we like.

For asymptotic stability with infinitesimal rate of decay, λ in condition (25) can be taken as infinitesimally small, and δ is zero. It can be seen that as λ and α tend to zero, conditions (25) and (28) approach each other. Thus, the result obtained through the original Lyapunov–Razumikhin theorem is a special case of the result obtained through the modified theorem.

IV. CONCLUSIONS

A method of applying the Lyapunov–Razumikhin sufficient condition for stability to feedback systems governed by ordinary nonlinear differential equations is presented. The results obtained show that the method is useful in that it is applicable to some systems which cannot be analyzed by other known methods.

A modified form of the Lyapunov–Razumikhin condition is presented, which establishes rate of decay of unforced response. Inputs can also be considered and BIBO stability studied.

REFERENCES

- [1] H. D'Angels, *Linear Time Varying Systems*. Boston: Allyn & Bacon, 1970.
- [2] V. M. Popov, "Absolute stability of nonlinear systems of automatic control," in *Nonlinear Systems Stability Analysis*, Aggerwal and Vidyasagar, Eds. Dowden, Hutchinson & Ross, 1977.
- [3] J. K. Hale, "Functional differential equations," in *Applied Mathematical Sciences*, vol. 3. New York: Springer-Verlag, 1971.
- [4] G. Seifert, "Liapunov–Razumikhin condition for asymptotic stability in functional differential equation of volterra type," *J. Diff. Equations*, vol. 16, no. 2, pp. 289–299, 1974.

Properties of Optimal Weighted Sensitivity Designs

Kathryn E. Lenz

Abstract—Weighted sensitivity designs for a class of single-input/single-output, linear, time-invariant systems are shown to have the property that the optimal controller has fewer right-half plane poles than the plant has right-half plane zeros. Weighted complementary sensitivity designs are shown to have the property that the optimal controller has fewer right-half plane zeros than the plant has right-half plane poles. Effects of various choices of weighting functions on the optimal solutions are described.

1. INTRODUCTION

Using a result due to Poreda [7], we derive basic frequency domain qualitative properties of the optimal solutions to weighted sensitivity minimization and weighted complementary sensitivity minimization problems for single-input/single-output, linear, time-invariant systems. We show that under certain continuity assumptions applied to the plant and weighting function, weighted sensitivity optimal controllers have fewer right-half plane poles than the plant has right-half plane zeros. Conversely, if a given stabilizing controller has fewer right-half plane poles than the plant has right-half plane zeros and makes the magnitude of the weighted sensitivity function flat across frequencies, then the controller is optimal.

The analogous result for weighted complementary sensitivity designs is that the optimal controller has fewer right-half plane zeros than the plant has right-half plane poles. Conversely, if a given stabilizing controller has fewer right-half plane zeros than the plant has right-half plane poles and makes the magnitude of the weighted complementary sensitivity function flat across frequencies, then the controller is optimal.

Certain weighted sensitivity and weighted complementary sensitivity problems are not optimized by a stabilizing controller, even though the optimization is done over the set of all stabilizing controllers. We describe situations in which this can happen for plants with $j\mathbb{R}$ poles or zeros. We also show that if one wishes to obtain a stabilizing controller as the optimal solution to a weighted sensitivity or weighted complementary sensitivity optimization problem, one should not use a weighting function with a $j\mathbb{R}$ zero.

II. NOTATION, DEFINITIONS, AND ASSUMPTIONS

Denote the imaginary axis of the complex plane \mathbb{C} by $j\mathbb{R}$ and the imaginary axis together with infinity by $j\overline{\mathbb{R}}$. We say that s is in the right-half plane, \mathbb{C}^+ , if $\text{Re}(s) > 0$. Correspondingly, s is in the left-half plane, \mathbb{C}^- , if $\text{Re}(s) < 0$. The subset of functions in L^∞ on $j\mathbb{R}$ that are also continuous on $j\overline{\mathbb{R}}$ will be denoted by L_c^∞ . The space of functions that are analytic in \mathbb{C}^+ and extend to functions in L^∞ on $j\mathbb{R}$ will be denoted by H^∞ . The functions in H^∞ having continuous extensions from \mathbb{C}^+ to $\mathbb{C}^+ \cup j\overline{\mathbb{R}}$ will be denoted by A_∞ .

For M a function of the complex variable s , M^* will denote $\overline{M(-\bar{s})}$. A function $M \in H^\infty$ is inner if $M^*M = 1$ a.e. on $j\mathbb{R}$. A Blaschke product $B(s)$ of degree m is an inner function of the form $B(s) = k \prod_{j=1}^m (1 - \bar{z}_j s)/(1 + z_j s)$, where $|k| = 1$ and $z_j \in \mathbb{C}^+$ for $j = 1, 2, \dots, m$. For B a finite Blaschke product, $B^* = B^{-1}$.

Manuscript received March 12, 1993; revised May 15, 1994. This work was supported in part by the National Science Foundation Grant ECS-9116214.

The author is with the Department of Mathematics and Statistics, University of Minnesota at Duluth, Duluth, MN 55812 USA.

IEEE Log Number 9407261.

A function $N \in H^\infty$ is outer if the closure of NH^2 is H^2 , where H^2 is the space of functions that are square integrable on $j\mathbb{R}$ and analytic in \mathbb{C}^+ . A function $F \in H^\infty$ is a unit in H^∞ if $F^{-1} \in H^\infty$.

If there exist functions $X, U, V, D \in H^\infty$ such that $P = XD^{-1}$, and $U^*X + V^*D = I$, then we call XD^{-1} a coprime factorization of P . We say that K stabilizes P under negative feedback if the functions $(1 + PK)^{-1}$, $PK(1 + PK)^{-1}$, $K(1 + PK)^{-1}$, and $P(1 + PK)^{-1}$ are in H^∞ . If such a K exists for P , we say that P is stabilizable.

Suppose that C has a coprime factorization $C = U^*V^{-1}$ such that $U^*X + V^*D = I$, where $P = XD^{-1}$, and $X, U, V, D \in H^\infty$. Then C stabilizes P under negative feedback. The Youla parameterization of the set of functions that stabilize P under negative feedback, parameterized in terms of $C = U^*V^{-1}$, is given by

$$\Sigma(P) := \{K: K = (U - D\hat{Q})(V + X\hat{Q})^{-1}, \hat{Q} \in H^\infty \text{ and } \hat{Q} \neq -V^*X^{-1}\}.$$

Assumption 1: The plant $P(s)$ is stabilizable and has a coprime factorization $P = XD^{-1} = X_0X_\infty D^{-1}$, where X_0 is a finite Blaschke product, X_∞ is an outer function, and D is a rational function with inner/outer factorization $D = D_1D_\infty$.

Let a generalized zero of a Lebesgue measurable function f on a Lebesgue measurable subset of $j\mathbb{R}$ be, as defined in [3], a point on $j\mathbb{R}$ for which the magnitude of f is not essentially bounded from zero on any neighborhood of the point. If f is continuous at a generalized zero, the generalized zero is a zero of f . Denote the set of generalized zeros of f on $j\mathbb{R}$ by $Z(f)$.

For a function f continuous and nonvanishing on $j\mathbb{R}$ and meromorphic in \mathbb{C}^+ , the winding number (or index) of $f(j\omega)$ may be computed as $\text{wno}(f) = \gamma_f - p_f$, where γ_f and p_f denote the number of \mathbb{C}^+ zeros and \mathbb{C}^+ poles of f respectively, counted according to their multiplicities.

For a given controller C and plant P , denote the number of \mathbb{C}^+ poles of C by p_c , the number of \mathbb{C}^+ zeros of C by z_c , the number of \mathbb{C}^+ poles of P by p_p , and the number of \mathbb{C}^+ zeros of P by z_p , counted according to their multiplicities. We show in Section III that the optimality, as described in (3.3) and (3.5), of $C \in \Sigma(P)$ depends in part on the number of \mathbb{C}^+ poles and \mathbb{C}^+ zeros of P and C .

Lemma 2.1. If $g(s) \in A_\infty$ and is nonvanishing on $j\mathbb{R}$, then $g(s)$ has at most a finite number of \mathbb{C}^+ zeros.

Proof. Suppose that $g(s) \in A_\infty$ and is nonvanishing on $j\mathbb{R}$. Since $g(s)$ is analytic and not identically zero in \mathbb{C}^+ , each \mathbb{C}^+ zero of $g(s)$ has a finite order and the set of all zeros of $g(s)$ has no limit point in \mathbb{C}^+ . Suppose that $g(s)$ has an infinite number of \mathbb{C}^+ zeros. Let $\phi(s) = (s-1)/(s+1) = z$. The function $\phi(s)$ conformally maps \mathbb{C}^+ onto the open unit disk. The function $g\phi\omega^{-1}$ is analytic in the open unit disk and continuous and nonvanishing on the unit circle. Since g has an infinite number of zeros in \mathbb{C}^+ , $g\phi\omega^{-1}$ has an infinite number of zeros in the open unit disk. But this set of zeros has no limit point in the open unit disk. There must be a sequence of these zeros with a limit point on the unit circle. Therefore $g\phi\omega^{-1}$ has a zero on the unit circle. This implies that $g(s)$ has a zero on $j\mathbb{R}$, which is a contradiction.

III. QUALITATIVE PROPERTIES OF OPTIMAL SOLUTIONS

Both weighted sensitivity and weighted complementary sensitivity optimization problems can be transformed into the one-block general distance problem of finding an optimal norm γ and if possible an optimal $\hat{Q} \in H^\infty$ such that

$$\gamma = \inf_{Q \in H^\infty} \|R - Q\|_\infty = \|R - \hat{Q}\|_\infty. \quad (3.1)$$

When $R \in L^\infty$ the infimum in (3.1) is achieved for some $Q \in H^\infty$ [4, p. 135]. If $R \in L^\infty$, then the optimal \hat{Q} is unique and $|R - \hat{Q}|$ is constant a.e. on $j\mathbb{R}$ [5, p. 196].

From Poreda's Theorem 1 [7, p. 250] we have that for R continuous and nonvanishing on $j\mathbb{R}$

$$\inf_{Q \in H^\infty} \|R - Q\|_\infty = \|R\|_\infty \quad (3.2)$$

if and only if $|R|$ is constant on $j\mathbb{R}$ and $\text{wno}(R) < 0$. From Poreda's Corollary ([7, p. 250] see also [1, p. 52]), if R is also meromorphic in \mathbb{C}^+ , (3.2) holds if and only if $R = \gamma B_1^* B_2$, where B_1 and B_2 are Blaschke products with $\deg(B_1) > \deg(B_2)$ and γ is a constant.

Throughout this section, suppose that $P = XD^{-1}$ satisfies Assumption 1 and a controller C is given such that $C = U^*V^{-1}$ for $U, V \in H^\infty$, and $U^*X + V^*D = I$. Also assume that neither P nor C is the zero function.

Define the weighted sensitivity transfer function $\Lambda(s, K)$ to be $\Lambda(K) = [W_1/(1 + PK)]$. Assume that the weighting function W_1 is analytic in \mathbb{C}^+ and has no zeros in \mathbb{C}^+ . Assume that $W_1 D_0 X_\infty \in H^\infty$ has at most finitely many generalized zeros on $j\mathbb{R}$ and is continuous at each point of $Z(W_1 D_0 X_\infty)$. Further assume that $W_1 D_0 V \in A_\infty$.

Parameterize the set of all stabilizing controllers $\Sigma(P)$ in terms of C . Then the weighted sensitivity optimization problem associated with Λ consists of finding γ_1 and \hat{K} such that

$$\begin{aligned} \gamma_1 = \|\Lambda(K)\|_\infty &= \inf_{K \in \Sigma(P)} \left\| \frac{W_1}{1 + PK} \right\|_\infty \\ &= \inf_{Q \in H^\infty} \|W_1 D V + W_1 X \hat{Q}\|_\infty. \end{aligned} \quad (3.3)$$

Define

$$\gamma_2 = \inf_{Q \in H^\infty} \|X_0^* W_1 D_0 V - Q\|_\infty. \quad (3.4)$$

To equate the optimization problem (3.3) with the one block general distance problem (3.4) we rule out the possibility that $\gamma_2 < \gamma_1$. To do this assume that $\gamma_2 \geq |W_1 D_0 V(j\omega)|$ for every $j\omega \in Z(W_1 D_0 X_\infty)$. Then, [2, p. 12], $\gamma_2 = \gamma_1$. This assumption need only be checked when P has a generalized zero on $j\mathbb{R}$. With this assumption C is optimal for (3.3) if and only if $Q = 0$ optimizes the right-hand side of (3.4).

Let W_2 be analytic in \mathbb{C}^+ and have no zeros in \mathbb{C}^+ . Assume that $W_2 X_\infty D_0 \in A_\infty$ has at most finitely many zeros on $j\mathbb{R}$. Also assume that $W_2 X_0 U \in A_\infty$. Define the weighted complementary sensitivity transfer function $\Theta(s, K)$ to be $\Theta(K) = [W_2 PK/(1 + PK)]$.

The weighted complementary sensitivity optimization problem associated with Θ consists of finding γ_3 and \hat{K} such that

$$\begin{aligned} \gamma_3 = \|\Theta(K)\|_\infty &= \inf_{K \in \Sigma(P)} \left\| \frac{W_2 PK}{1 + PK} \right\|_\infty \\ &= \inf_{Q \in H^\infty} \|W_2 X U - W_2 X D Q\|_\infty. \end{aligned} \quad (3.5)$$

Define

$$\gamma_4 = \inf_{Q \in H^\infty} \|D_0^* W_2 X_0 U - Q\|_\infty. \quad (3.6)$$

To equate the optimization problem (3.5) with (3.6), assume that $\gamma_4 \geq |W_2 X_0 U(j\omega)|$ for every $j\omega \in Z(W_2 X_\infty D_0)$. Then, [2, p. 12], $\gamma_4 = \gamma_3$. This assumption need only be checked when P has a pole on $j\mathbb{R}$. With this assumption C is optimal for (3.5) if and only if $Q = 0$ optimizes the right-hand side of (3.6).

Theorem 3.7. A controller $C \in \Sigma(P)$ is uniquely optimal for the weighted sensitivity problem if and only if $|\Lambda(C)|$ is constant a.e. on $j\mathbb{R}$ and $p_c < z_p$.

Proof: Parameterizing $\Sigma(P)$ in terms of $C = U^*V^{-1}$ we have that C is uniquely optimal for (3.3) if and only if $Q = 0$ is uniquely optimal for (3.4). Assume that $Q = 0$ is optimal for (3.4). Then since $X_i^*W_1D_oV \in L^\infty$, $Q = 0$ is uniquely optimal and $|\Lambda(C)| = |X_i^*W_1D_oV|$ is constant on $j\bar{\mathbb{R}}$. Therefore $X_i^*W_1D_oV$ does not vanish anywhere on $j\bar{\mathbb{R}}$ and [7, p. 250], $\text{wno}(X_i^*W_1D_oV) < 0$. The product $X_i^*W_1D_oV$ is meromorphic in \mathcal{C}^+ since it can be expressed as the quotient of analytic functions in \mathcal{C}^+ . Since X_i is a finite Blaschke product and $X_i^*W_1D_oV$ does not vanish on $j\bar{\mathbb{R}}$, W_1D_oV does not vanish on $j\bar{\mathbb{R}}$. Then by Lemma 2.1, W_1D_oV has a finite number of \mathcal{C}^+ zeros. The \mathcal{C}^+ zeros of W_1D_oV are the \mathcal{C}^+ poles of C and the \mathcal{C}^+ zeros of X_i are the \mathcal{C}^+ zeros of P . Therefore $\text{wno}(X_i^*W_1D_oV) = p_i - z_p$. Thus $z_p > p_i$.

Conversely, suppose that $z_p > p_i$ and $|\Lambda(C)|$ is constant a.e. on $j\bar{\mathbb{R}}$. Then $|X_i^*W_1D_oV|$ is constant on $j\bar{\mathbb{R}}$. Since $X_i^*W_1D_oV$ is continuous and not identically zero on $j\bar{\mathbb{R}}$, it does not vanish anywhere on $j\bar{\mathbb{R}}$. Then $\text{wno}(X_i^*W_1D_oV) = p_i - z_p < 0$. Therefore [7, p. 250], $Q = 0$ is uniquely optimal.

Theorem 3.8: A controller $C \in \Sigma(P)$ is uniquely optimal for the weighted complementary sensitivity problem if and only if $|\Theta(C)|$ is constant a.e. on $j\bar{\mathbb{R}}$ and $z_i < p_p$.

Proof: Parameterizing $\Sigma(P)$ in terms of $C = U^*V^{-1}$, we have that C is uniquely optimal for (3.5) if and only if $Q = 0$ is uniquely optimal for (3.6). Assume that $Q = 0$ is optimal for (3.6). Since $D_i^*W_2X_oU \in L^\infty$, $Q = 0$ is uniquely optimal and $|\Theta(C)| = |D_i^*W_2X_oU|$ is constant on $j\bar{\mathbb{R}}$. Therefore $D_i^*W_2X_oU$ does not vanish anywhere on $j\bar{\mathbb{R}}$ and $\text{wno}(D_i^*W_2X_oU) < 0$. The product $D_i^*W_2X_oU$ is meromorphic in \mathcal{C}^+ since it can be expressed as the quotient of analytic functions in \mathcal{C}^+ . Since D_i is a finite Blaschke product and $D_i^*W_2X_oU$ does not vanish on $j\bar{\mathbb{R}}$, W_2X_oU does not vanish on $j\bar{\mathbb{R}}$. Then by Lemma 2.1, W_2X_oU has a finite number of \mathcal{C}^+ zeros. The \mathcal{C}^+ zeros of W_2X_oU are the \mathcal{C}^+ zeros of U and the \mathcal{C}^+ zeros of D_i are the \mathcal{C}^+ poles of P . Therefore $\text{wno}(D_i^*W_2X_oU) = z_i - p_p$. Thus $p_p > z_i$.

Conversely, suppose that $p_p > z_i$ and $|\Theta(C)|$ is constant a.e. on $j\bar{\mathbb{R}}$. Then $|D_i^*W_2X_oU|$ is constant on $j\bar{\mathbb{R}}$. Since $D_i^*W_2X_oU$ is assumed continuous and not identically zero on $j\bar{\mathbb{R}}$, it is nonzero on $j\bar{\mathbb{R}}$. Then $\text{wno}(D_i^*W_2X_oU) = z_i - p_p < 0$. Therefore $Q = 0$ is uniquely optimal.

Example 3.9: Let sequences $\{\gamma_n\}$ and $\{\phi_n\}$ consist of the positive numbers satisfying $\cos \gamma_n \sinh \gamma_n = \sin \gamma_n \cosh \gamma_n$, and $\cos \phi_n \cosh \phi_n = 1$. Assume the ordering $\phi_j < \phi_k$ and $\gamma_j < \gamma_k$ for $j < k$. Let $P = X_o D_i^{-1}$, where $D_i = (s - 1)/(s + 1)$, and

$$X_o = \frac{1}{(s+1)} \prod_{n=1}^{\infty} \frac{1 + \epsilon s + s^2 \gamma_n^{-1}}{1 + \epsilon s + s^2 \phi_n^{-1}}.$$

One can show that X_o belongs to A_∞ and is outer [6]. Let $U(s)$ be the constant $X_o(1)^{-1}$. Then $V = (1 - U^*X_o)D_i^{-1}$ belongs to H^∞ and $U^*X_o + V^*D_i = 1$. So $C = U^*V^{-1} \in \Sigma(P)$. Define the weighting function $W_2 = (U^*X_o)^{-1}$. Since U^*X_o has no poles or zeros in \mathcal{C}^+ , W_2 is analytic in \mathcal{C}^+ and has no zeros in \mathcal{C}^+ . Also $|\Theta(C)| = |W_2X_oU| = 1$ a.e. on $j\bar{\mathbb{R}}$. Since $p_p = 1$, and $z_i = 0$, by Theorem 3.8 $C = U^*V^{-1}$ is optimal for (3.5).

Remark 3.10: Theorem 3.7 requires that P have at least one \mathcal{C}^+ zero. Suppose that $z_p = 0$ and that $P = X_o D_i^{-1}$ with D_i inner and X_o a unit in H^∞ . Let $C_k = kX_o^{-1}$. For $k > 1$, $C_k \in \Sigma(P)$ and $W_1/(1 + PC_k) = W_1 D_i/(D_i + k)$. For each $j\omega \in j\bar{\mathbb{R}}$

$$\lim_{k \rightarrow \infty} \left| \frac{W_1}{1 + PC_k} \right|^2 = \lim_{k \rightarrow \infty} \frac{W_1^* W_1}{1 + k D_i + k D_i^* + k^2} = 0.$$

Thus $\inf_{K \in \Sigma(P)} \|\Lambda(K)\|_\infty = 0$. No $K \in \Sigma(P)$ achieves this minimum.

Remark 3.11: Theorem 3.8 requires that P have at least one \mathcal{C}^+ pole. When $p_p = 0$ and P has no poles on $j\bar{\mathbb{R}}$, $K = 0$ is optimal for the weighted complementary sensitivity problem. One can check that $0 \in \Sigma(P)$ and that $\|\Theta(0)\|_\infty = 0$.

IV. WEIGHTING FUNCTIONS

In Section III we observed that the optimal controller makes the weighted sensitivity or weighted complementary sensitivity function flat across frequencies. In this section we explore the implications of this for various choices of weighting functions. Suppose throughout this section that the plant P and given controller C satisfy the same assumptions as in Section III. Let $U = U^*U_o$ and $V = V^*V_o$ be inner/outer factorizations.

One might expect that a controller that optimizes a weighted sensitivity problem would have poor complementary sensitivity properties. This is not necessarily the case.

Corollary 4.1: Suppose that $C' \in \Sigma(P)$ is the unique optimal controller for a weighted sensitivity problem. Suppose that C' has no generalized zeros on $j\bar{\mathbb{R}}$. Then C' is also uniquely optimal for the weighted complementary sensitivity problem with $W_2 = \alpha(X_o U_o)^{-1}$, where α is any nonzero constant, if and only if $z_i < p_p$.

Proof: This result follows from Theorem 3.8 after we observe that for $W_2 = \alpha(X_o U_o)^{-1}$, $|\Theta(C')| = |D_i^*W_2X_oU| = |\alpha|$ on $j\bar{\mathbb{R}}$.

Corollary 4.2: Suppose that $C' \in \Sigma(P)$ is the unique optimal controller for a weighted complementary sensitivity problem. Suppose that V_o has no generalized zeros on $j\bar{\mathbb{R}}$. Then C' is also uniquely optimal for the weighted sensitivity problem with $W_1 = \nu(D_o V_o)^{-1}$, where ν is any nonzero constant, if and only if $p_i < z_p$.

Proof: Observe that $|\Lambda(C')| = |X_i^*W_1D_oV| = |\nu|$ on $j\bar{\mathbb{R}}$. Apply Theorem 3.7.

In weighted sensitivity optimization the weighting function and the plant determine the optimal Nyquist plot of $PC'(j\omega)$ versus ω . In the case of a constant weight we can say precisely what the optimal Nyquist plot is.

Corollary 4.3: Suppose that $W_1 = \nu$, for some constant $\nu > 0$, $C' \in \Sigma(P)$, and $p_i < z_p$. Then C' is optimal for the weighted sensitivity problem if and only if for each $\omega \in j\bar{\mathbb{R}}$, $PC'(j\omega)$ satisfies the equation of a circle with center $c_r = (-1, 0)$ and radius $r = \nu/z_i$, where γ is the optimal norm in (3.3). Furthermore, $r < 1$.

Proof: By Theorem 3.7, C' is optimal for (3.4) if and only if $|\Lambda(C')|$ is constant on $j\bar{\mathbb{R}}$. Let $PC'(j\omega) = x + jy$. Then for each $j\omega \in j\bar{\mathbb{R}}$, $|\Lambda(C')| = \gamma$ if and only if $\nu^2/(1/(1+x)^2 + y^2)) = \gamma^2$. Equivalently, $y^2 + (1+x)^2 = \nu^2/\gamma^2$. From the Maximum Modulus Theorem, since $\Lambda(C')$ is analytic in \mathcal{C}^+ and $\Lambda(C') = \nu$ at every \mathcal{C}^+ zero of P , $\|\Lambda(C')\|_\infty \geq \nu$. Since $\Lambda(C')$ is nonconstant in \mathcal{C}^+ , $\|\Lambda(C')\|_\infty > \nu$.

Corollary 4.4: Suppose that $W_2 = \mu$, for some constant $\mu > 0$, $C' \in \Sigma(P)$, and $z_i < p_p$. Then C' is optimal for the weighted complementary sensitivity problem if and only if for each $\omega \in j\bar{\mathbb{R}}$, $PC'(j\omega)$ satisfies the equation of a circle with center $c_r = (c_{rx}, 0)$ and radius $r = \gamma\mu/(\gamma^2 - \mu^2)$, where $c_{rx} = -\gamma^2/(\gamma^2 - \mu^2)$, where $\gamma > \mu$ is the optimal norm in (3.5).

Proof: By Theorem 3.8, C' is uniquely optimal for (3.5) if and only if $|\Theta(C')| = \gamma$ a.e. on $j\bar{\mathbb{R}}$ for some constant γ . From the Maximum Modulus Theorem, since $\Theta(C')$ is analytic in \mathcal{C}^+ and $\Theta(C') = \mu$ at every \mathcal{C}^+ pole of P , $\gamma \geq \mu$. Since $\Theta(C')$ is nonconstant in \mathcal{C}^+ , $\gamma > \mu$. Let $PC'(j\omega) = x + jy$. Then for each $j\omega \in j\bar{\mathbb{R}}$, $|\Lambda(C')| = \gamma$ if and only if $\mu^2((x^2 + y^2)/(1+x)^2 + y^2)) = \gamma^2$. Equivalently, $y^2 + (x + \gamma^2/(\gamma^2 - \mu^2))^2 = \gamma^2\mu^2/(\gamma^2 - \mu^2)^2$.

Certain weighted sensitivity and weighted complementary sensitivity problems are not optimized by a stabilizing controller, even though the optimization is done over the set of all stabilizing controllers.

We describe situations in which this happens in Corollary 4.5 and Corollary 4.6

Corollary 4.5 Suppose that $W_1 D$ has a zero on $j\bar{\mathbb{R}}$. Then $C \in \Sigma(P)$ is not optimal for the weighted sensitivity problem.

Proof Suppose that $C \in \Sigma(P)$ is optimal for (3.3). Then by Theorem 3.7 $|\Lambda(C)|$ is constant a.e. on $j\bar{\mathbb{R}}$. Since $W_1 D$ has a zero at $j\omega_c \in j\bar{\mathbb{R}}$, then $|\Lambda(C)| = 0$ a.e. on $j\bar{\mathbb{R}}$. But no $K \in \Sigma(P)$ can achieve this.

Thus when a plant P has a $j\bar{\mathbb{R}}$ pole with multiplicity n , no stabilizing controller will minimize the weighted sensitivity function if W_1 does not also have the pole with multiplicity at least n . For example consider $P = (s+1)/s$. For $\Lambda = 1$ and $D = s/(s+1)$, ΛD^{-1} is a coprime factorization of P . For any W_1 that does not have a pole at $s = 0$, problem (3.3) will not be optimized by a stabilizing controller.

Also, Corollary 4.5 shows that if one wishes to obtain a stabilizing controller as the optimal solution to a weighted sensitivity problem, one should not choose W_1 strictly proper or with a $j\bar{\mathbb{R}}$ zero.

Corollary 4.6 Suppose that $p_1 > 0$ and $W_1 \Lambda$ has a zero on $j\bar{\mathbb{R}}$. Then $C \in \Sigma(P)$ is not optimal for the weighted complementary sensitivity problem.

Proof Suppose that $C \in \Sigma(P)$ is optimal for (3.5). By Theorem 3.8 $|\Theta(C)|$ is constant a.e. on $j\bar{\mathbb{R}}$. Since $W_1 \Lambda$ has a zero at $j\omega_c \in j\bar{\mathbb{R}}$, $|\Theta(C)| = 0$ a.e. on $j\bar{\mathbb{R}}$. But this contradicts the fact that since $p_1 > 0$ by the Maximum Modulus theorem $\|\Theta(K)\|_\infty > 0$ for all $K \in \Sigma(P)$.

Thus when an unstable plant P has a $j\bar{\mathbb{R}}$ zero, no stabilizing controller will minimize the weighted complementary sensitivity function if $W_1 P$ also has the zero. Corollary 4.6 also shows that if one wishes to obtain a stabilizing controller as the optimal solution to a weighted complementary sensitivity optimization problem with an unstable plant, one should not use a weighting function with a $j\bar{\mathbb{R}}$ zero.

ACKNOWLEDGMENT

The author thanks Prof. A. Tannenbaum for helpful discussions regarding this work.

REFERENCES

- [1] V. M. Adamjan, D. Z. Arov, and M. G. Krein, "Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem," *Math. USSR Sb.*, vol. 15, pp. 31-73, 1971.
- [2] D. S. Flamm, "Outer factor absorption for H^∞ control problems," Dep. Electrical Engineering, Princeton Univ., Princeton, NJ, ISS Report 55, 1990.
- [3] D. S. Flamm and H. Yang, "Optimal mixed sensitivity for SISO distributed plants," *IEEE Trans. Automat. Contr.*, vol. 39, no. 6, pp. 1150-1165, 1994.
- [4] J. B. Garnett, *Bounded Analytic Functions*. New York: Academic, 1981.
- [5] P. Koosis, *Introduction to H^1 Spaces*. Cambridge: Cambridge Univ. Press, 1980.
- [6] K. Lenz, H. Özbay, A. Tannenbaum, I. Tutu, and B. Morton, "Frequency domain analysis and robust control design for an ideal flexible beam," *Automatica*, vol. 27, pp. 947-961, 1991.
- [7] S. J. Poreda, "A characterization of badly approximable functions," *Trans. American Math. Soc.*, vol. 169, pp. 249-256, 1972.

A Periodic Fixed-Structure Approach to Multirate Control

Wassim M. Haddad and Vikram Kapila

Abstract—In this note we develop an approach to designing reduced-order multirate controllers. A discrete-time model that accounts for the multirate timing sequence of measurements is presented and is shown to have periodically time-varying dynamics. Using discrete-time stability theory, the optimal projection approach to fixed-order (i.e., full- and reduced-order) dynamic compensation is generalized to obtain reduced-order periodic controllers that account for the multirate architecture. It is shown that the optimal reduced-order controller is characterized by means of a periodically time-varying system of equations consisting of coupled Riccati and Lyapunov equations. In addition, the multirate static output-feedback control problem is considered. For both problems, the design equations are presented in a concise, unified manner to facilitate their accessibility for developing numerical algorithms for practical applications.

I. INTRODUCTION

Many applications of feedback control involve continuous time systems subject to digital (discrete time) control. Furthermore, in practical applications, the control system actuators and sensors have differing bandwidths. For example, in flexible structure control, it is not unusual to attenuate the low frequency high amplitude modes by means of low bandwidth actuators that are relatively heavy and hence able to exert high force/torque to control the higher frequency modes. Obviously, the high bandwidth actuators would require sensors that are sampled at high rates, while low bandwidth actuators require only sensors sampled at low data rates. As a consequence, the use of various sensor data rates leads to a multirate control problem. To properly use such data, a multirate controller must carefully account for the timing sequence of incoming data. The purpose of this note is to develop a general approach to full- and reduced order steady state multirate dynamic compensation.

Multirate control problems have been of interest for many years with increased emphasis in recent years [1]-[3], [9]-[11], [15]. A common feature of these papers is the realization that the multirate sampling process leads to periodically time varying dynamics. Hence, with a suitable reinterpretation, results on multirate control can also be applied to single rate or multirate problems involving systems with periodically time varying dynamics. The principal challenge of these problems is to arrive at a tractable control design formulation in spite of the extreme complexity of such systems. In order to account for the periodic time varying dynamics of multirate systems, a periodically time varying control law architecture was proposed in [10] and [11] which appears promising in this regard. An alternative approach which has been proposed for the multirate control problem is the use of an expanded state space formulation [2]. However, this approach results in very high order systems and is often numerically intractable. Finally, a cost translation and a lifting approach to the multirate LQG problem has been proposed in [15] which does not lead to an increase in the state dimension. Specifically, [15] shows how to translate a multirate sampled data LQG problem into an equivalent modified single rate shift invariant problem via a lifted isomorphism.

Manuscript received May 10, 1993; revised May 15, 1994 and July 15, 1994. This work was supported in part by the National Science Foundation under Grants ECS 9109558 and ECS 9350181.

The authors are with the School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0150 USA.

IEEE Log Number 9406991.

However, this approach results in an equivalent system involving more inputs and outputs than the original system. The interested reader is referred to [7], [8], [10], and [14] for further discussions on multirate and periodic control.

For generality in our development, we consider both full- and reduced-order dynamic compensators as well as static output-feedback controllers. In the discrete-time case, this problem was considered in [4], while single rate sampled-data aspects were addressed in [5]. The approach of the present note is the fixed-structure Riccati equation technique developed in [4]. Essentially, this approach addresses controller complexity by explicitly imposing implementation constraints on the controller structure, and optimizing over that class of controllers. Specifically, in addressing the problem of reduced-order dynamic compensation, it is shown in [4] that optimal reduced-order steady-state dynamic compensators can be characterized by means of an algebraic system of Riccati/Lyapunov equations coupled by a projection matrix which arises as a direct consequence of optimality and which represents a breakdown of the separation between the operations of state estimation and state estimate feedback; that is, the certainty equivalence principle is no longer valid. The proof is based on expressing the closed-loop quadratic cost functional as a function of the design parameters, i.e., the compensator gains, and the utilization of Lagrange multipliers for optimization over the parameter space. Thus, this approach provides a constrained optimal control methodology in which we do not seek to optimize a performance measure per se, but rather a performance measure within a class of *a priori* fixed-structure controllers.

In the present note, analog-to-digital conversions are employed within a multirate setting to obtain periodically time-varying dynamics. The compensator is thus assigned a corresponding discrete-time periodic structure to account for the multirate measurements. It is shown that the optimal reduced-order multirate dynamic compensator is characterized by a periodically time-varying system of four equations consisting of two modified Riccati equations and two modified Lyapunov equations corresponding to each intermediate point of the periodicity interval. Because of the time-varying nature of the problem, the necessary conditions for optimality now involve multiple projections corresponding to each intermediate point of the periodic interval and whose rank along the periodic interval is equal to the order of the compensator. Similar extensions to reduced-order multirate estimation are addressed in [12].

Nomenclature

$I_r, 0, \infty, 0,$	$r \times r$ identity matrix, $r \times s$ zero matrix, $0, \infty,$
$(\cdot)^T, (\cdot)^{-1}, \text{tr}(\cdot), \mathcal{E}$	transpose, inverse, trace, expected value
$\mathcal{R}, \mathcal{R}^{*s}, \mathcal{R}^t$	real numbers, $r \times s$ real matrices, $\mathcal{R}^{t \times t}$
k, α	discrete-time indices
n, m, l_k, n_c, \hat{n}	positive integers; $1 \leq n_c \leq n; \hat{n} = n + n_c$
x, y, x_c, u	$n-, l_k-, n_c-, m$ -dimensional vectors
$A, B, C(k)$	$n \times n, n \times m, l_k \times n$ matrices
$A_c(k), B_c(k), C_c(k), D_c(k)$	$n_c \times n_c, n_c \times l_k, m \times n_c, m \times l_k$ matrices
R_1, R_2	$n \times n, m \times m$ state and control weightings; $R_1 \geq 0, R_2 > 0$
R_{12}	$n \times m$ cross weighting; $R_1 - R_{12}R_2^{-1}R_{12}^T \geq 0$

II. THE FIXED-STRUCTURE MULTIRATE STATIC AND DYNAMIC CONTROL PROBLEMS

In this section, we state the fixed-structure static and dynamic, sampled-data, multirate output-feedback control problems. In the problem formulation, the sample intervals h_k and dynamic compensator order n_c are fixed, and the optimization is performed over the compensator parameters $[A_c(\cdot), B_c(\cdot), C_c(\cdot), D_c(\cdot)]$. For design trade-off studies, h_k and n_c can be varied and the problem can be solved for each pair of values of interest.

Fixed-Structure Multirate Static Output-Feedback Control Problem: Consider the n th-order continuous-time system with multirate sampled-data measurements

$$\dot{x}(t) = Ax(t) + Bu(t) + w_1(t), \quad t \in [0, \infty), \quad (2.1)$$

$$y(t_k) = C(t_k)x(t_k) + w_2(t_k), \quad k = 1, 2, \dots \quad (2.2)$$

Then design a static output-feedback multirate sampled-data control law

$$u(t_k) = D_c(t_k)y(t_k) \quad (2.3)$$

which, with D/A zero-order-hold controls $u(t) = u(t_k), t \in [t_k, t_{k+1})$, minimizes the quadratic performance criterion

$$\mathcal{J}_s(D_c(\cdot)) \triangleq \lim_{t \rightarrow \infty} \mathcal{E} \frac{1}{t} \int_0^t [x^T(s)R_1x(s) + 2x^T(s)R_{12}u(s) + u^T(s)R_2u(s)] ds. \quad (2.4)$$

Fixed-Structure Multirate Dynamic Output-Feedback Control Problem: Given the n th-order continuous-time system (2.1) with multirate sampled-data measurements (2.2), design an n_c th-order ($1 \leq n_c \leq n$) multirate sampled-data dynamic compensator

$$x_c(t_{k+1}) = A_c(t_k)x_c(t_k) + B_c(t_k)y(t_k) \quad (2.5)$$

$$u(t_k) = C_c(t_k)x_c(t_k) + D_c(t_k)y(t_k) \quad (2.6)$$

which, with D/A zero-order-hold controls $u(t) = u(t_k), t \in [t_k, t_{k+1})$, minimizes the quadratic performance criterion (2.4) with $\mathcal{J}_s(D_c(\cdot))$ denoted by $\mathcal{J}_d(A_c(\cdot), B_c(\cdot), C_c(\cdot), D_c(\cdot))$.

The key feature of both problems is the time-varying nature of the output equation (2.2) which represents sensor measurements available at different rates. Fig. 1 provides a typical multirate timing diagram for a three-sensor model. For generality, we do not assume that the sample intervals $h_k \triangleq t_{k+1} - t_k$ are uniform (note the sample times for sensor #3 in Fig. 1). However, we do assume that the overall timing sequence of intervals $[t_k, t_{k+N}]$, $k = 1, 2, \dots$, is periodic over $[0, \infty)$, where N represents the periodic interval. Note that $h_{k+N} = h_k, k = 1, 2, \dots$. Since different sensor measurements are available at different times t_k , the dimension l_k of the measurements $y(t_k)$ may also vary periodically. Finally, in subsequent analysis, the static output-feedback law (2.3) and dynamic compensator (2.5) and (2.6) are assigned periodic gains corresponding to the periodic timing sequence of the multirate measurements.

In the above problem formulation, $w_1(t)$ denotes a continuous-time stationary white noise process with nonnegative-definite intensity $V_1 \in \mathcal{R}^{n \times n}$, while $w_2(t_k)$ denotes a variable-dimension discrete-time white noise process with positive-definite covariance $V_2(t_k) \in \mathcal{R}^{l_k \times l_k}$. We assume $w_2(t_k)$ is cyclostationary, that is, $V_2(t_{k+N}) = V_2(t_k), k = 1, 2, \dots$.

In what follows, we shall simplify the notation considerably by replacing the continuous-time sample instant t_k by the discrete-time index k . With this minor abuse of notation, we replace $x(t_k)$ by $x(k)$, $x_c(t_k)$ by $x_c(k)$, $y(t_k)$ by $y(k)$, $u(t_k)$ by $u(k)$, $w_2(t_k)$ by $w_2(k)$, $A_c(t_k)$ by $A_c(k)$ [and similarly for $B_c(\cdot), C_c(\cdot)$, and

$D_c(\cdot)$, $C(t_k)$ by $C(k)$, and $V_2(t_k)$ by $V_2(k)$. The context should clarify whether the argument is " k " or " t_k ." With this notation, our periodicity assumption on the compensator implies $A_c(k+N) = A_c(k)$, $k = 1, 2, \dots$, and similarly for $B_c(\cdot)$, $C_c(\cdot)$, and $D_c(\cdot)$. Also, by assumption, $C(k+N) = C(k)$, for $k = 1, 2, \dots$.

Next, we model the propagation of the plant over one time step. For notational convenience, define $H(k) \triangleq \int_0^{h_k} e^{A_s} ds$.

Theorem 2.1: For the fixed-order multirate sampled-data control problem, the plant dynamics (2.1) and quadratic performance criterion (2.4) have the equivalent discrete-time representation

$$x(k+1) = A(k)x(k) + B(k)u(k) + w_1'(k), \quad (2.7)$$

$$y(k) = C(k)x(k) + w_2(k), \quad (2.8)$$

$$\begin{aligned} J = \delta_\infty + \lim_{K \rightarrow \infty} \frac{1}{K} \mathcal{E} \sum_{k=1}^K [x^T(k) R_1(k) x(k) \\ + 2x^T(k) R_{12}(k) u(k) + u^T(k) R_2(k) u(k)] \end{aligned} \quad (2.9)$$

where

$$A(k) \triangleq e^{A h_k}, \quad B(k) \triangleq H(k) B,$$

$$w_1'(k) \triangleq \int_0^{h_k} e^{A(h_k-s)} w_1(k+s) ds,$$

$$\delta_\infty \triangleq \lim_{K \rightarrow \infty} \frac{1}{K} \text{tr} \sum_{k=1}^K \frac{1}{h_k} \int_0^{h_k} \int_0^s e^{A^T V_1} e^{A^T r} R_1 dr ds,$$

$$R_1(k) \triangleq \frac{1}{h_k} \int_0^{h_k} e^{A^T s} R_1 e^{A s} ds,$$

$$R_{12}(k) \triangleq \frac{1}{h_k} \int_0^{h_k} e^{A^T s} R_1 H(s) B ds + \frac{1}{h_k} H^T(k) R_{12},$$

$$\begin{aligned} R_2(k) \triangleq R_2 + \frac{1}{h_k} \int_0^{h_k} [B^T H^T(s) R_1 H(s) B \\ + R_{12}^T H(s) B + B^T H^T(s) R_{12}] ds \end{aligned}$$

and $w_1'(k)$ is a zero-mean discrete-time white noise process with

$$\mathcal{E}\{w_1'(k) w_1'^T(k)\} = V_1(k) \triangleq \int_0^{h_k} e^{A^T s} V_1 e^{A^T s} ds.$$

The proof of this theorem is a straightforward calculation involving integrals of white noise signals, and hence is omitted. Note that by the sampling periodicity assumption, $A(k+N) = A(k)$, $k = 1, 2, \dots$.

The above formulation assumes that a discrete-time multirate measurement model is available. One can assume, alternatively, that analog measurements corrupted by continuous-time white noise are available instead, that is, $y(t) = Cx(t) + w_2(t)$. In this case, one can develop an equivalent discrete-time model that employs an averaging-type A/D device [5] $\hat{y}(k) = (1/h_k) \int_{t_k}^{t_{k+1}} y(t) dt$. It can be shown that the resulting averaged measurements depend upon delayed samples of the state. In this case, the equivalent discrete-time model can be captured by a suitably augmented system. For details see [5].

Remark 2.1: The equivalent discrete-time quadratic performance criterion (2.9) involves a constant offset δ_∞ ¹ which is a function of sampling rates, and effectively imposes a lower bound on sampled-data performance due to the discretization process.

¹As will be shown by Lemma 3.1, due to the periodicity of h_k , δ_∞ is a constant.

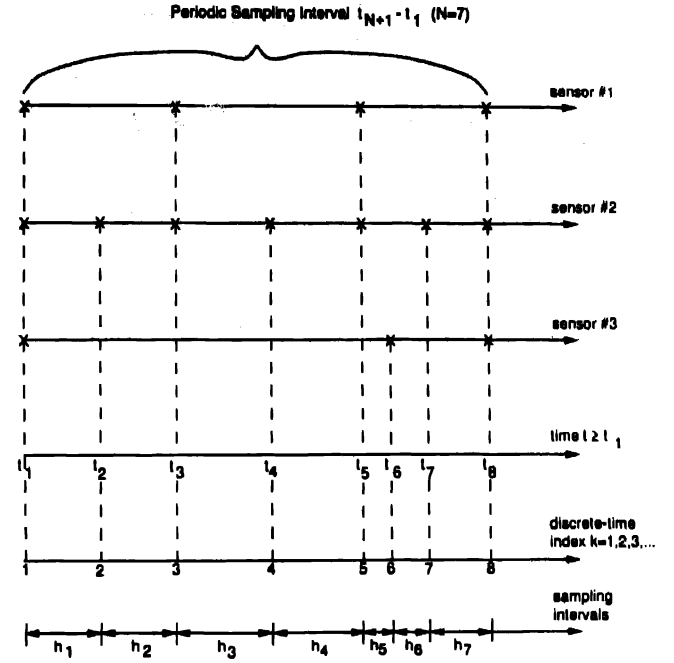


Fig. 1. Multirate timing diagram for sampled-data control

III. THE FIXED-STRUCTURE MULTIRATE SAMPLED-DATA STATIC OUTPUT-FEEDBACK PROBLEM

In this section, we obtain necessary conditions that characterize solutions to the multirate sampled-data static output-feedback control problem. First, we form the closed-loop system for (2.7), (2.8), and (2.3) to obtain

$$x(k+1) = \hat{A}(k)x(k) + \hat{w}(k) \quad (3.1)$$

where $\hat{A}(k) \triangleq A(k) + B(k)D_c(k)C(k)$. The closed-loop disturbance $\hat{w}(k) \triangleq w_1'(k) + B(k)D_c(k)w_2(k)$, $k = 1, 2, \dots$, has nonnegative-definite covariance $\hat{V}(k) \triangleq V_1(k) + B(k)D_c(k)V_2(k)D_c^T(k)B^T(k)$, where we assume that the noise correlation $V_{12}(k) \triangleq \mathcal{E}\{w_1'(k)w_2^T(k)\} = 0$, that is, the continuous-time plant noise and the discrete-time measurement noise are uncorrelated. The cost functional (2.9) can now be expressed in terms of the closed-loop second-moment matrix. The following result is immediate.

Proposition 3.1: For given $D_c(\cdot)$, the second-moment matrix $Q(k) \triangleq \mathcal{E}[x(k)x^T(k)]$, satisfies

$$Q(k+1) = \hat{A}(k)Q(k)\hat{A}^T(k) + \hat{V}(k). \quad (3.2)$$

Furthermore,

$$\begin{aligned} J_s(D_c(\cdot)) = \delta_\infty + \lim_{K \rightarrow \infty} \frac{1}{K} \text{tr} \\ \sum_{k=1}^K [Q(k)\hat{R}(k) + D_c^T(k)R_2(k)D_c(k)V_2(k)] \end{aligned} \quad (3.3)$$

where $\hat{R}(k) \triangleq R_1(k) + R_{12}(k)D_c(k)C(k) + C^T(k)D_c^T(k)R_{12}^T(k) + C^T(k)D_c^T(k)R_2(k)D_c(k)C(k)$.

Remark 3.1: Equation (3.2) is a periodic Lyapunov equation which has been extensively studied in [7] and [8].

We now show that the covariance Lyapunov equation (3.2) reaches a steady-state periodic trajectory as $K \rightarrow \infty$. For the next result, we introduce the parameterization, $k = \alpha + \beta N$, where the index α satisfies $1 \leq \alpha \leq N$, and $\beta = 1, 2, \dots$. We now restrict our

attention to output-feedback controllers having the property that the closed-loop transition matrix over one period $\hat{\Phi}_p(\alpha) \triangleq \hat{A}(\alpha + N - 1)\hat{A}(\alpha + N - 2) \cdots \hat{A}(\alpha)$, is stable for $\alpha = 1, \dots, N$. Note that since $\hat{A}(\cdot)$ is periodic, the eigenvalues of $\hat{\Phi}_p(\alpha)$ are actually independent of α . Hence, it suffices to require that $\hat{\Phi}_p(1) = \hat{A}(N)\hat{A}(N-1) \cdots \hat{A}(1)$ is stable.

Lemma 3.1 [12]: Suppose $\hat{\Phi}_p(1)$ is stable. Then for given $D_c(k)$, the covariance Lyapunov equation (3.2) reaches a steady-state periodic trajectory as $k \rightarrow \infty$, that is,

$$\lim_{k \rightarrow \infty} [Q(k), Q(k+1), \dots, Q(k+N-1)] \\ = [Q(\alpha), Q(\alpha+1), \dots, Q(\alpha+N-1)]. \quad (3.4)$$

In this case, the covariance $Q(k)$ satisfies

$$Q(\alpha+1) = \hat{A}(\alpha)Q(\alpha)\hat{A}^T(\alpha) + \hat{V}(\alpha), \quad \alpha = 1, \dots, N \quad (3.5)$$

where $Q(N+1) = Q(1)$. Furthermore, the quadratic performance criterion (3.3) is given by

$$\mathcal{J}_s(D_c(\cdot)) = \delta + \frac{1}{N} \text{tr} \sum_{\alpha=1}^N [Q(\alpha)\hat{R}(\alpha) + D_c^T(\alpha)R_2(\alpha)D_c(\alpha)V_2(\alpha)] \quad (3.6)$$

where

$$\delta \triangleq \frac{1}{N} \text{tr} \sum_{\alpha=1}^N \frac{1}{h_\alpha} \int_0^{h_\alpha} \int_0^s e^{A_1 \tau} V_1 \tau^{-1/2} R_1 d\tau ds.$$

For the statement of the main result of this section, define the set

$$S_s \triangleq \{D_c(\cdot) : \hat{\Phi}_p(\alpha) \text{ is stable, for } \alpha = 1, \dots, N\}. \quad (3.7)$$

In addition to ensuring that the covariance Lyapunov equation (3.2) reaches a steady-state periodic trajectory as $k \rightarrow \infty$, the set S_s constitutes sufficient conditions under which the Lagrange multiplier technique is applicable to the fixed-order multirate sampled-data static output-feedback control problem. The asymptotic stability of the transition matrix $\hat{\Phi}_p(\alpha)$ serves as a normality condition which further implies that the dual $P(\alpha)$ of $Q(\alpha)$ is nonnegative-definite.

For notational convenience in stating the multirate sampled-data static output-feedback result, define the notation

$$R_{2a}(\alpha) \triangleq B^T(\alpha)P(\alpha+1)B(\alpha) + \frac{1}{N}R_2(\alpha),$$

$$V_{2a}(\alpha) \triangleq C^T(\alpha)Q(\alpha)C^T(\alpha) + V_2(\alpha),$$

$$P_a(\alpha) \triangleq B^T(\alpha)P(\alpha+1)A(\alpha) + \frac{1}{N}R_{12}^T(\alpha),$$

$$Q_a(\alpha) \triangleq A(\alpha)Q(\alpha)C^T(\alpha)$$

for arbitrary $Q(\alpha)$ and $P(\alpha) \in \mathcal{R}^{n \times n}$ and $\alpha = 1, \dots, N$.

Theorem 3.1: Suppose $D_c(\cdot) \in S_s$ solves the multirate sampled-data static output-feedback control problem. Then there exist $n \times n$ nonnegative-definite matrices $Q(\alpha)$ and $P(\alpha)$ such that, for $\alpha = 1, \dots, N$, $D_c(\alpha)$ is given by

$$D_c(\alpha) = R_{2a}^{-1}(\alpha)P_a(\alpha)Q(\alpha)C^T(\alpha)V_{2a}^{-1}(\alpha) \quad (3.8)$$

and such that $Q(\alpha)$ and $P(\alpha)$ satisfy

$$Q(\alpha+1) = A(\alpha)Q(\alpha)A^T(\alpha) + V_1(\alpha) - Q_a(\alpha)V_{2a}^{-1}(\alpha)Q_a^T(\alpha) \\ + [Q_a(\alpha) + B(\alpha)D_c(\alpha)V_{2a}(\alpha)] \\ \cdot V_{2a}^{-1}(\alpha)[Q_a(\alpha) + B(\alpha)D_c(\alpha)V_{2a}(\alpha)]^T, \quad (3.9)$$

$$P(\alpha) = A^T(\alpha)P(\alpha+1)A(\alpha) \\ + \frac{1}{N}R_1(\alpha) - P_a^T(\alpha)R_{2a}^{-1}(\alpha)P_a(\alpha) \\ + [P_a(\alpha) + R_{2a}(\alpha)D_c(\alpha)C^T(\alpha)]^T \\ \cdot R_{2a}^{-1}(\alpha)[P_a(\alpha) + R_{2a}(\alpha)D_c(\alpha)C^T(\alpha)]. \quad (3.10)$$

Furthermore, the minimal cost is given by

$$\mathcal{J}_s(D_c(\cdot)) = \delta + \frac{1}{N} \text{tr} \sum_{\alpha=1}^N Q(\alpha)[R_1(\alpha) - 2R_{12}(\alpha)R_{2a}^{-1}(\alpha) \\ \cdot P_a(\alpha)Q(\alpha)C^T(\alpha)V_{2a}^{-1}(\alpha)C^T(\alpha) \\ + P_a^T(\alpha)R_{2a}^{-1}(\alpha)R_{2a}(\alpha)R_{2a}^{-1}(\alpha) \\ \cdot P_a(\alpha)Q(\alpha)C^T(\alpha)V_{2a}^{-1}(\alpha)C^T(\alpha)]. \quad (3.11)$$

Proof. The necessary conditions for optimality follow from standard Lagrange multiplier arguments. See [14] for a similar proof. \square

Remark 3.2. In the full-state feedback case, we take $C^T(\alpha) = I$, $V_2(\alpha) = 0$, and $R_{12}(\alpha) = 0$ for $\alpha = 1, \dots, N$. In this case, (3.8) becomes $D_c(\alpha) = -R_{2a}^{-1}(\alpha)B^T(\alpha)P(\alpha+1)A(\alpha)$, and (3.10) specializes to

$$P(\alpha) = A^T(\alpha)P(\alpha+1)A(\alpha) + \frac{1}{N}R_1(\alpha) \\ - A^T(\alpha)P(\alpha+1)B(\alpha)R_{2a}^{-1}(\alpha)B^T(\alpha)P(\alpha+1)A(\alpha) \quad (3.12)$$

while (3.9) is superfluous and can be omitted. Finally, we note that if we assume a single rate architecture, the plant dynamics are constant and (3.12) collapses to the standard discrete-time regulator Riccati equation.

IV. THE FIXED-STRUCTURE MULTIRATE SAMPLED-DATA DYNAMIC OUTPUT-FEEDBACK CONTROL PROBLEM

In this section, we consider the fixed order multirate sampled-data dynamic compensation problem. As in Section III, we first form the closed-loop system for (2.7), (2.8), (2.5), and (2.6) to obtain

$$x(k+1) = A(k)\bar{x}(k) + w(k) \quad (4.1)$$

where

$$\bar{x}(k) \triangleq \begin{bmatrix} x(k) \\ x_c(k) \end{bmatrix},$$

$$\bar{A}(k) \triangleq \begin{bmatrix} A(k) + B(k)D_c(k)C^T(k) & B(k)C^T(k) \\ B_c(k)C^T(k) & A_c(k) \end{bmatrix},$$

$$\bar{A}(k+N) = \bar{A}(k), \quad k = 1, 2, \dots$$

The closed-loop disturbance

$$w(k) \triangleq \begin{bmatrix} w_1^T(k) + B(k)D_c(k)w_2(k) \\ B_c(k)w_2(k) \end{bmatrix}, \quad k = 1, 2, \dots$$

has nonnegative-definite covariance shown by (x) at the bottom of the next page, where once again we assume that the continuous-time plant noise and the discrete-time measurement noise are uncorrelated, i.e., $V_{12}(k) \triangleq \mathcal{E}[w_1^T(k)w_2^T(k)] = 0$. As for the static output-feedback case, the cost functional (2.9) can now be expressed in terms of the closed-loop second-moment matrix. Specifically, Proposition 3.1 and Lemma 3.1 hold for the dynamic-output feedback problem with $x(k)$, $\bar{A}(k)$, $\bar{V}(k)$, and $\bar{R}(k)$ replaced by $x(k)$, $\bar{A}(k)$, $\bar{V}(k)$, and $\bar{R}(k)$, respectively, where $\bar{R}(k)$ is shown in (y) at the bottom of the next page.

For the next result, define the closed-loop transition matrix and the compensator transition matrix over one period by $\hat{\Phi}_p(\alpha) \triangleq \bar{A}(\alpha +$

$N-1) \bar{A}(\alpha+N-2) \cdots \bar{A}(\alpha)$ and $\Phi_{cp}(\alpha) \triangleq A_c(\alpha+N-1)A_c(\alpha+N-2) \cdots A_c(\alpha)$, respectively. Note that since $A_c(\alpha)$ is required to be periodic, the eigenvalues of $\Phi_{cp}(\alpha)$ are actually independent of α .

In the following, we obtain necessary conditions that characterize solutions to the fixed-order multirate sampled-data dynamic compensation problem. Derivation of these conditions requires additional technical assumptions. Specifically, we further restrict $[A_c(\cdot), B_c(\cdot), C_c(\cdot), D_c(\cdot)]$ to the set

$$S_c \triangleq \{(A_c(\alpha), B_c(\alpha), C_c(\alpha), D_c(\alpha)) : \Phi_p(\alpha) \text{ is stable and } (\Phi_{cp}(\alpha), B_{cp}(\alpha), C_{cp}(\alpha)) \text{ is controllable and observable, } \alpha = 1, \dots, N\} \quad (4.2)$$

where

$$B_{cp}(\alpha) \triangleq [A_c(\alpha+N-1)A_c(\alpha+N-2) \cdots A_c(\alpha+1)B_c(\alpha), A_c(\alpha+N-1)A_c(\alpha+N-2) \cdots A_c(\alpha+2)B_c(\alpha+1), \dots, B_c(\alpha+N-1)]. \quad (4.3)$$

$$C_{cp}(\alpha) \triangleq \begin{bmatrix} C_c(\alpha+N-1)A_c(\alpha+N-2) \cdots A_c(\alpha) \\ C_c(\alpha+N-2)A_c(\alpha+N-3) \cdots A_c(\alpha) \\ \vdots \\ C_c(\alpha) \end{bmatrix}. \quad (4.4)$$

The set S_c constitutes sufficient conditions under which the Lagrange multiplier technique is applicable to the fixed-order multirate sampled-data control problem. This is similar to concepts involving *moving equilibrium* for periodic Lyapunov/Riccati equations discussed in [7] and [8]. Specifically, the formulas for the lifted isomorphism (4.3) and (4.4) are equivalent to assuming the stability of $\bar{A}(\cdot)$ along with the reachability and observability of $[A_c(\cdot), B_c(\cdot), C_c(\cdot)]$ [8], [10]. The asymptotic stability of the transition matrix $\Phi_p(\alpha)$ serves as a normality condition which further implies that the dual $\hat{P}(\alpha)$ of $\hat{Q}(\alpha) \triangleq \mathcal{E}[\tilde{x}(\alpha)\tilde{x}^T(\alpha)]$ is nonnegative-definite. Furthermore, the assumption that $[\Phi_{cp}(\alpha), B_{cp}(\alpha), C_{cp}(\alpha)]$ is controllable and observable is a nondegeneracy condition which implies that the lower right $n_c \times n_c$ subblocks of $\hat{Q}(\alpha)$ and $\hat{P}(\alpha)$ are positive definite, thus yielding explicit gain expressions for $A_c(\alpha)$, $B_c(\alpha)$, $C_c(\alpha)$, and $D_c(\alpha)$.

In order to state the main result, we require some additional notation and a lemma concerning pairs of nonnegative-definite matrices. For details see [6].

Lemma 4.1: Let \hat{Q} , \hat{P} be $n \times n$ nonnegative-definite matrices and assume $\text{rank } \hat{Q}\hat{P} = n_c$. Then there exist $n_c \times n$ matrices G , Γ and an $n_c \times n_c$ invertible matrix M , unique except for a change of basis in \mathbb{R}^{n_c} , such that $\hat{Q}\hat{P} = G^T M \Gamma$, $\Gamma G^T = I_{n_c}$. Furthermore, the $n \times n$ matrices $\tau \triangleq G^T \Gamma$, $\tau_\perp \triangleq I_n - \tau$, are idempotent and have $\text{rank } n_c$ and $n - n_c$, respectively.

The following result gives necessary conditions that characterize solutions to the fixed-order multirate sampled-data control problem. For convenience in stating this result, recall the definitions of $R_{2a}(\cdot)$, $V_{2a}(\cdot)$, $P_a(\cdot)$, and $Q_a(\cdot)$ and define the additional notation

$$M(\alpha) \triangleq \begin{bmatrix} I_n \\ D_c(\alpha)C_c(\alpha) \end{bmatrix}, \quad \hat{M}(\alpha) \triangleq \begin{bmatrix} I_n \\ -R_{2a}^{-1}(\alpha)P_a(\alpha) \end{bmatrix}.$$

$$\hat{R}(\alpha) \triangleq \begin{bmatrix} R_1(\alpha) & R_{12}(\alpha) \\ R_{12}^T(\alpha) & R_2(\alpha) \end{bmatrix}$$

for arbitrary $P(\alpha) \in \mathbb{R}^{n \times n}$ and $\alpha = 1, \dots, N$.

Theorem 4.1: Suppose $[A_c(\cdot), B_c(\cdot), C_c(\cdot), D_c(\cdot)] \in S_c$ solves the fixed-order multirate sampled-data dynamic output-feedback control problem. Then there exist $n \times n$ nonnegative-definite matrices $\hat{Q}(\alpha)$, $\hat{P}(\alpha)$, $\hat{Q}(\alpha)$, and $\hat{P}(\alpha)$ such that, for $\alpha = 1, \dots, N$, $A_c(\alpha)$, $B_c(\alpha)$, $C_c(\alpha)$, and $D_c(\alpha)$ are given by

$$A_c(\alpha) = \Gamma(\alpha+1)[A(\alpha) - B(\alpha)R_{2a}^{-1}(\alpha)P_a(\alpha) - Q_a(\alpha)V_{2a}^{-1}(\alpha)C_c(\alpha) - B(\alpha)D_c(\alpha)C_c(\alpha)]G^T(\alpha), \quad (4.5)$$

$$B_c(\alpha) = \Gamma(\alpha+1)[Q_a(\alpha)V_{2a}^{-1}(\alpha) + B(\alpha)D_c(\alpha)], \quad (4.6)$$

$$C_c(\alpha) = -[R_{2a}^{-1}(\alpha)P_a(\alpha) + D_c(\alpha)C_c(\alpha)]G^T(\alpha), \quad (4.7)$$

$$D_c(\alpha) = -R_{2a}^{-1}(\alpha)P_a(\alpha)Q_c(\alpha)C_c^T(\alpha)V_{2a}^{-1}(\alpha) \quad (4.8)$$

and such that $\hat{Q}(\alpha)$, $\hat{P}(\alpha)$, $\hat{Q}(\alpha)$, and $\hat{P}(\alpha)$ satisfy

$$\begin{aligned} \hat{Q}(\alpha+1) = & A(\alpha)Q(\alpha)A^T(\alpha) + V_1(\alpha) - Q_a(\alpha)V_{2a}^{-1}(\alpha)Q_a^T(\alpha) \\ & + \tau_\perp(\alpha+1)\{[A(\alpha) - B(\alpha)R_{2a}^{-1}(\alpha)P_a(\alpha)] \\ & \cdot \hat{Q}(\alpha)\{A(\alpha) - B(\alpha)R_{2a}^{-1}(\alpha)P_a(\alpha)\}^T \\ & + \{Q_a(\alpha) + B(\alpha)D_c(\alpha)V_{2a}(\alpha)\}V_{2a}^{-1}(\alpha) \\ & \cdot \{Q_a(\alpha) + B(\alpha)D_c(\alpha)V_{2a}(\alpha)\}^T\}\tau_\perp(\alpha+1), \end{aligned} \quad (4.9)$$

$$\begin{aligned} P(\alpha) = & A^T(\alpha)P(\alpha+1)A(\alpha) + \frac{1}{N}R_1(\alpha) \\ & - P_a^T(\alpha)R_{2a}^{-1}(\alpha)P_a(\alpha) \\ & + \tau_\perp^T(\alpha)\{[A(\alpha) - Q_a(\alpha)V_{2a}^{-1}(\alpha)C_c(\alpha)]^T \hat{P}(\alpha+1) \\ & \cdot \{A(\alpha) - Q_a(\alpha)V_{2a}^{-1}(\alpha)C_c(\alpha)\} \\ & + \{P_a(\alpha) + R_{2a}(\alpha)D_c(\alpha)C_c(\alpha)\}^T R_{2a}^{-1}(\alpha) \\ & \cdot \{P_a(\alpha) + R_{2a}(\alpha)D_c(\alpha)C_c(\alpha)\}\}\tau_\perp(\alpha), \end{aligned} \quad (4.10)$$

$$\begin{aligned} \hat{Q}(\alpha+1) = & \tau(\alpha+1)\{[A(\alpha) - B(\alpha)R_{2a}^{-1}(\alpha)P_a(\alpha)]\hat{Q}(\alpha) \\ & \cdot \{A(\alpha) - B(\alpha)R_{2a}^{-1}(\alpha)P_a(\alpha)\}^T \\ & + \{Q_a(\alpha) + B(\alpha)D_c(\alpha)V_{2a}(\alpha)\}V_{2a}^{-1}(\alpha) \\ & \cdot \{Q_a(\alpha) + B(\alpha)D_c(\alpha)V_{2a}(\alpha)\}^T\}\tau^T(\alpha+1), \end{aligned} \quad (4.11)$$

$$\begin{aligned} \hat{P}(\alpha) = & \tau^T(\alpha)\{[A(\alpha) - Q_a(\alpha)V_{2a}^{-1}(\alpha)C_c(\alpha)]^T \hat{P}(\alpha+1) \\ & \cdot \{A(\alpha) - Q_a(\alpha)V_{2a}^{-1}(\alpha)C_c(\alpha)\} \\ & + \{P_a(\alpha) + R_{2a}(\alpha)D_c(\alpha)C_c(\alpha)\}^T R_{2a}^{-1}(\alpha) \\ & \cdot \{P_a(\alpha) + R_{2a}(\alpha)D_c(\alpha)C_c(\alpha)\}\}\tau(\alpha) \end{aligned} \quad (4.12)$$

$$\text{rank } \hat{Q}(\alpha) = \text{rank } \hat{P}(\alpha) = \text{rank } \hat{Q}(\alpha)\hat{P}(\alpha) = n_c. \quad (4.13)$$

$$\hat{V}(k) \triangleq \begin{bmatrix} V_1(k) + B(k)D_c(k)V_2(k)D_c^T(k)B^T(k) & B(k)D_c(k)V_2(k)B_c^T(k) \\ B_c(k)V_2(k)D_c^T(k)B^T(k) & B_c(k)V_2(k)B_c^T(k) \end{bmatrix} \quad (x)$$

$$\hat{R}(k) \triangleq \begin{bmatrix} R_1(k) + R_{12}(k)D_c(k)C_c(k) + C_c^T(k)D_c^T(k)R_{12}^T(k) & R_{12}(k)C_c(k) + C_c^T(k)D_c^T(k)R_2(k)C_c(k) \\ + C_c^T(k)D_c^T(k)R_2(k)D_c(k)C_c(k) & C_c^T(k)R_{12}^T(k) + C_c^T(k)R_2(k)D_c(k)C_c(k) \\ C_c^T(k)R_{12}^T(k) + C_c^T(k)R_2(k)D_c(k)C_c(k) & C_c^T(k)R_2(k)C_c(k) \end{bmatrix}. \quad (y)$$

Furthermore, the minimal cost is given by

$$\mathcal{J}_c(A_c(\cdot), B_c(\cdot), C_c(\cdot), D_c(\cdot)) = \delta + \frac{1}{N} \text{tr} \sum_{\alpha=1}^N [\{M(\alpha)Q(\alpha)M^T(\alpha) + \hat{M}(\alpha)\hat{Q}(\alpha)\hat{M}^T(\alpha)\}\hat{R}(\alpha)]. \quad (4.14)$$

Proof: The proof is identical to the proof of Theorem 3.3 of [14] concerning reduced-order control of discrete-time linear periodic systems with suitable reinterpretations for capturing the equivalent multirate sampled-data model. \square

Theorem 4.1 provides necessary conditions for the fixed-order multirate sampled-data control problem. These necessary conditions consist of a system of two modified periodic difference Lyapunov equations and two modified periodic difference Riccati equations coupled by projection matrices $\tau(\alpha)$, $\alpha = 1, \dots, N$. As expected, these equations are periodically time-varying over the period $1 \leq \alpha \leq N$ in accordance with the multirate nature of the measurements. As discussed in [4], the fixed-order constraint on the compensator gives rise to the projection τ which characterizes the optimal reduced-order compensator gains. In the multirate case, however, it is interesting to note that the time-varying nature of the problem gives rise to multiple projections corresponding to each of the intermediate points of the periodicity interval, and whose rank along the periodic interval is equal to the order of the compensator.

Remark 4.3: As in the linear time-invariant case [4] to obtain the full-order multirate LQG controller, set $n_c = n$. In this case, the projections $\tau(\alpha)$, and $\Gamma(\alpha)$ and $G(\alpha)$, for $\alpha = 1, \dots, N$, become the identity. Consequently, (4.11) and (4.12) play no role and hence can be omitted. In order to draw connections with existing full-order multirate results, set $D_c(\alpha) = 0$ and $R_{12}(\alpha) = 0$, $\alpha = 1, \dots, N$, so that

$$A_c(\alpha) = A(\alpha) - B(\alpha)R_{2\alpha}^{-1}(\alpha)B^T(\alpha)P(\alpha+1)A(\alpha) - A(\alpha)Q(\alpha)C^T(\alpha)V_{2\alpha}^{-1}(\alpha)C^T(\alpha), \quad (4.15)$$

$$B_c(\alpha) = A(\alpha)Q(\alpha)C^T(\alpha)V_{2\alpha}^{-1}(\alpha), \quad (4.16)$$

$$C_c(\alpha) = -R_{2\alpha}^{-1}(\alpha)B^T(\alpha)P(\alpha+1)A(\alpha) \quad (4.17)$$

where $Q(\alpha)$ and $P(\alpha)$ satisfy

$$Q(\alpha+1) = A(\alpha)Q(\alpha)A^T(\alpha) + V_1(\alpha) - A(\alpha)Q(\alpha)C^T(\alpha)V_{2\alpha}^{-1}(\alpha)C^T(\alpha)Q(\alpha)A^T(\alpha), \quad (4.18)$$

$$P(\alpha) = A^T(\alpha)P(\alpha+1)A(\alpha) + \frac{1}{N}R_1(\alpha) - A^T(\alpha)P(\alpha+1)B(\alpha)R_{2\alpha}^{-1}(\alpha)B^T(\alpha)P(\alpha+1)A(\alpha). \quad (4.19)$$

Thus, the full-order multirate sampled-data controller is characterized by two decoupled periodic difference Riccati equations (observer and regulator Riccati equations) over the period $\alpha = 1, \dots, N$. This corresponds to the results obtained in [10]. Next, assuming a single rate architecture yields time-invariant plant dynamics, while (4.18) and (4.19) specialize to the discrete-time observer and regulator Riccati equations. Alternatively, retaining the reduced-order constraint and assuming single rate sampling, Theorem 4.1 yields the sampled-data optimal projection equations for reduced-order dynamic compensation given in [5].

V. NUMERICAL EVALUATION OF INTEGRALS INVOLVING MATRIX EXPONENTIALS

To evaluate the integrals involving matrix exponentials appearing in Theorem 2.1, we utilize the approach of [16]. The idea is to eliminate the need for integration by computing the matrix exponential of appropriate block matrices.

Proposition 5.1: For $\alpha = 1, \dots, N$, consider the following partitioned matrix exponentials

$$\begin{bmatrix} E_1 & E_2 & E_3 & E_4 \\ 0_n & E_5 & E_6 & E_7 \\ 0_n & 0_n & E_8 & E_9 \\ 0_{m \times n} & 0_{m \times n} & 0_{m \times n} & I_m \end{bmatrix} \triangleq \exp \begin{bmatrix} -A^T & I_n & 0_n & 0_{n \times m} \\ 0_n & -A^T & R_1 & 0_{n \times m} \\ 0_n & 0_n & A & B \\ 0_{m \times n} & 0_{m \times n} & 0_{m \times n} & 0_m \end{bmatrix} h_\alpha$$

$$\begin{bmatrix} E_{10} & E_{11} & E_{12} & E_{13} \\ 0_n & E_{14} & E_{15} & E_{16} \\ 0_n & 0_n & E_{17} & E_{18} \\ 0_{m \times n} & 0_{m \times n} & 0_{m \times n} & I_m \end{bmatrix} \triangleq \exp \begin{bmatrix} -A^T & I_n & 0_n & 0_{n \times m} \\ 0_n & -A^T & R_1 & R_{12} \\ 0_n & 0_n & A & B \\ 0_{m \times n} & 0_{m \times n} & 0_{m \times n} & 0_m \end{bmatrix} h_\alpha$$

$$\begin{bmatrix} E_{19} & E_{20} & E_{21} \\ 0_n & E_{22} & E_{23} \\ 0_n & 0_n & E_{24} \end{bmatrix} \triangleq \exp \begin{bmatrix} -A & I_n & 0_n \\ 0_n & -A & V_1 \\ 0_n & 0_n & A^T \end{bmatrix} h_\alpha$$

of orders $(3n+m) \times (3n+m)$, $(3n+m) \times (3n+m)$, and $3n \times 3n$, respectively. Then, for $\alpha = 1, \dots, N$,

$$A(\alpha) = E_{21}^T, \quad B(\alpha) = E_{18}, \quad V_1(\alpha) = E_{21}^T E_{21},$$

$$R_1(\alpha) = \frac{1}{h_\alpha} E_{17}^T E_{15}, \quad R_{12}(\alpha) = \frac{1}{h_\alpha} E_{17}^T E_{16},$$

$$R_2(\alpha) = R_2 + \frac{1}{h_\alpha} [B^T E_{17}^T E_{15} + E_{15}^T E_{17} B - B^T E_{17}^T E_{15}].$$

$$\delta = \frac{1}{N} \sum_{\alpha=1}^N \frac{1}{h_\alpha} \text{tr} R_1 E_{21}^T E_{21}.$$

The proof of the above proposition involves straightforward manipulations of matrix exponentials.

VI. ILLUSTRATIVE NUMERICAL EXAMPLE

For illustrative purposes, we consider a numerical example involving a rigid body with a flexible appendage. This example is reminiscent of a single-axis spacecraft involving unstable dynamics and sensor fusion of slow, accurate spacecraft attitude sensors (such as horizon sensors or star trackers) with fast, less accurate rate gyroscopes. The motivation for slow/fast sensor configuration is that rate information can be used to improve the attitude control between attitude measurements. Hence define

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & -0.01 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

$$C = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 0.1 & 0 & 0.1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}^T.$$

$$V_1 = DD^T, \quad V_2 = I_2, \quad E = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$R_1 = E^T E, \quad R_2 = 1.$$

Note that the dynamic model involves one rigid body mode along with one flexible mode at a frequency of 1 rad/s with 0.5% damping. The matrix C captures the fact that the rigid body angular position and tip velocity of the flexible appendage are measured. Also, note that the rigid body position measurement is corrupted by the flexible mode (i.e., observation spillover). To reflect a plausible mission, we assume that the rigid body angular position is measured by an attitude sensor sampling at 1 Hz, while the tip appendage velocity is measured by a rate gyro sensor sampling at 5 Hz. The matrix R_1 expresses the desire to regulate the rigid body and tip appendage positions, and the matrix V_1 was chosen to capture the type of noise correlation that arises when the dynamics are transformed into a modal basis.

Using the homotopy algorithm based on a prediction and a Newton correction scheme for periodic difference Riccati equations reported in [13], the following designs were obtained. For $n_c = 4$ discrete-time single rate and multirate controllers were obtained from (4.15)–(4.19) using Theorem 2.1 for continuous-time to discrete-time conversions. These designs were compared using the performance criterion (4.14). The results are summarized as follows

Measurement Scheme	Optimal Cost
Two 1 Hz sensors	65.6922
Two 5 Hz sensors	53.9930
Multirate scheme (1 Hz and 5 Hz sensors)	54.8061

Note that the improvement in the cost of the two 5 Hz sensor scheme over the multirate scheme is minimal, which clearly demonstrates that the multirate scheme provides sensor complexity reduction over the two 5 Hz sensor scheme.

REFERENCES

- [1] R. Aracil, A. J. Avella, and V. Felio, "Multirate sampling technique in digital control systems simulation," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-14, pp. 776–780, 1984.
- [2] M. Araki and K. Yamamoto, "Multivariable multirate sampled-data systems: State-space description, transfer characteristics and Nyquist criterion," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 145–154, 1986.
- [3] M. C. Berg, N. Amit, and D. Powell, "Multirate digital control system design," *IEEE Trans. Automat. Contr.*, vol. AC-33, pp. 1139–1150, 1988.
- [4] D. S. Bernstein, L. D. Davis, and D. C. Hyland, "The optimal projection equations for reduced-order discrete-time modeling, estimation and control," *AIAA J. Guidance, Contr. Dynamics*, vol. 9, pp. 288–293, 1986.
- [5] D. S. Bernstein, L. D. Davis, and S. W. Grealey, "The optimal projection equations for fixed-order, sampled-data dynamic compensation with computational delay," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 859–862, 1986.
- [6] D. S. Bernstein and W. M. Haddad, "Robust stability and performance via fixed-order dynamic compensation with guaranteed cost bounds," *Math. Contr. Signals, Syst.*, vol. 3, pp. 139–163, 1990.
- [7] S. Bittanti, P. Colaneri, and G. DeNicolao, "The difference periodic Riccati equation for the periodic prediction problem," *IEEE Trans. Automat. Contr.*, vol. AC-33, pp. 706–712, 1988.
- [8] P. Bolzern and P. Colaneri, "The periodic Lyapunov equation," *SIAM J. Matrix Anal. Appl.*, vol. 9, pp. 499–512, 1988.
- [9] J. R. Broussard and N. Haylo, "Optimal multirate output feedback," in *Proc. IEEE Conf. Dec. Contr.*, Las Vegas, NV, 1984, pp. 926–929.
- [10] P. Colaneri, R. Scattolini, and N. Schiavoni, "LQG optimal control of multirate sampled-data systems," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 675–682, 1992.

- [11] D. P. Glasson, "A new technique for multirate digital control," *AIAA J. Guidance, Contr., Dynamics*, vol. 5, pp. 379–382, 1982.
- [12] W. M. Haddad, D. S. Bernstein, and V. Kapila, "Reduced-order multirate estimation," *AIAA J. Guidance, Contr., Dynamics*, vol. 17, pp. 712–721, 1994.
- [13] W. M. Haddad and V. Kapila, "A periodic fixed-structure approach to multirate control," in *Proc. IEEE Conf. Dec. Contr.*, San Antonio, TX, 1993, pp. 1791–1796.
- [14] W. M. Haddad, V. Kapila, and E. G. Collins, Jr., "Optimality conditions for reduced-order modeling, estimation, and control for discrete-time linear periodic plants," *J. Math. Syst., Estim., Contr.*, to appear.
- [15] D. G. Meyer, "Cost translation and a lifting approach to the multirate LQG problem," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1411–1415, 1992.
- [16] C. F. Van Loan, "Computing integrals involving the matrix exponential," *IEEE Trans. Automat. Contr.*, vol. AC-23, pp. 395–404, 1978.

Pursuing a Maneuvering Target Which Uses a Random Process for Its Control

V. E. Beneš, K. L. Helmes, and R. W. Rishel

Abstract—Since a pursuer pursuing a maneuvering target does not know what maneuvers an evading target will make, the maneuvers (the target's control law) appear as a random process to the pursuer. However, he has opinions about what the evader will do. From these, he can assign a prior probability distribution to the evader's maneuvers.

For a linear pursuit evasion problem in which the evader's control law is modeled as a random process, in which the pursuer has partial noisy linear measurements of his own and the evader's relative position, and a quadratic optimality criterion is used, past results of the authors imply that the optimal control is a linear function of the "predicted miss." Determining the predicted miss involves estimating the evader's terminal position from past system measurements. Nonlinear filtering techniques are used to give expressions for computing the conditional expectation of the evader's terminal position even in the presence of the random unknown maneuvers of the evader.

1. INTRODUCTION

Recently, Helmes and Rishel [5], [6], and Beneš [1] have shown, for a linear system,

$$dx_t = (Ax_t + Bu_t)dt + dz_t \quad (1)$$

whose driving noises z_t may not be Gaussian, and for an optimality criteria which is a quadratic function of the terminal state plus a quadratic running cost in the control, that the optimal control is a linear function of a quantity called the "predicted miss."

The purpose of this note is to apply these results, and results from optimal nonlinear filtering and extrapolation, to describe the optimal control for a linear pursuit-evasion problem in which the evader's control is modeled as an unknown random process.

The main problem in computing the predicted miss is to compute the conditional expectation of the evader's terminal position given

Manuscript received April 30, 1993; revised February 15, 1994. This work was supported in part by the National Science Foundation under Grant DMS-9105649.

V. E. Beneš is with The Beneš Group, 26 Taylor St., Millburn, NJ 07041, USA.

K. L. Helmes and R. W. Rishel are with the Department of Mathematics, University of Kentucky, Lexington, KY 40506 USA.

IEEE Log Number 9406992.

where

$$\alpha_t \triangleq e^{\int_0^t (H y_s^\#)' dm_s - (1/2) \int_0^t |H y_s^\#|^2 ds} \quad (45)$$

and

$$\beta_t = e^{\int_0^t (H y_s^\nu)' (dm_s - H y_s^\# ds) - (1/2) \int_0^t |H y_s^\nu|^2 ds}. \quad (46)$$

Thus, from (36),

$$E[y_t | \mathcal{M}_t] = \frac{E_0[\hat{y}_t | \mathcal{M}_t]}{E_0[\hat{\Lambda}_t | \mathcal{M}_t]} = \frac{E_0[y_s^\# \alpha_t \beta_t | \mathcal{M}_t]}{E_0[\alpha_t \beta_t | \mathcal{M}_t]} + \frac{E_0[y_t^\nu \alpha_t \beta_t | \mathcal{M}_t]}{E_0[\alpha_t \beta_t | \mathcal{M}_t]}. \quad (47)$$

Now, from (40) and (45), we see that $y_t^\#$ and α_t are \mathcal{M}_t measurable. Thus, they may be taken out of the conditional expectations in (47) giving

$$E[y_t | \mathcal{M}_t] = y_t^\# + \frac{E_0[y_t^\nu \beta_t | \mathcal{M}_t]}{E_0[\beta_t | \mathcal{M}_t]}. \quad (48)$$

Now, again, since under P_0 , v_t and m_t are independent processes, the conditional expectations in (48) may be replaced by ordinary expectations with respect to the distribution of v_t giving (38). A similar argument gives (39). That β_t is a solution of (42) follows by differentiating (46) using Ito's rule.

Remark 1: Formulas (21) and (22) [or (38) and (39)] express the conditional expectations we are interested in, in terms of the expected value with respect to the distribution of the process v_t of quantities which are solutions of stochastic differential equations driven by the measurements m_t . It is not surprising since we do not know v_t that we should have to take an expected value with respect to its distribution. However, evaluating this expected value in the general case could be prohibitively complex. For instance, to determine the quantities whose expected value is to be taken in (38) and (39), equations (40)–(42) must be solved. There is one equation—(41)—for each control path and the solution of this equation is a coefficient in (42) for β_t . Thus, (42) could be an infinite system of equations, one for each control path, which are driven by the incoming measurements m_t .

The difficulties mentioned in Remark 1 disappear in the following case.

Finiteness Assumption: The probability distribution of the control v_t is concentrated on a finite number of paths

$$v_t^1, v_t^2, \dots, v_t^n \quad (49)$$

and is given by

$$P[v_t = v_t^i] = P^i, \quad i = 1, \dots, n. \quad (50)$$

Theorem II, in this case, is easily seen to reduce to the following theorem.

Theorem 3: If the finiteness assumption holds, then

$$E[y_t | \mathcal{M}_t] = y_t^\# + \frac{\sum_{i=1}^n P^i y_t^i \beta_t^i}{\sum_{i=1}^n P^i \beta_t^i} \quad (51)$$

and

$$E[v_t | \mathcal{M}_t] = \frac{\sum_{i=1}^n P^i v_t^i \beta_t^i}{\sum_{i=1}^n P^i \beta_t^i} \quad (52)$$

where $y_t^\#$ is the solution of (40), y_t^i is the solution of

$$dy_t^i = [(C' - R(t)H'H)y_t^i + Dv_t^i] dt; \quad y_0^i = 0 \quad (53)$$

and β_t^i is the solution of

$$d\beta_t^i = \beta_t^i (H y_t^i)' (dm_t - H y_t^\# dt); \quad \beta_0^i = 1. \quad (54)$$

Remark 2: In Theorem 3, the solutions of the linear system (53) driven by the n known paths v_t^i could be precomputed to give the n functions y_t^i . Thus, (40) for $y_t^\#$ and (54) for the n functions β_t^i are a system of $n+1$ equations driven by the measurements m_t which must be solved to determine the quantities in (51) and (52). This is an implementable method for computing these conditional expectations.

Remark 3: In the case of a general distribution for the paths of the control process v_t , the following could be considered as a numerical method for determining the conditional distributions in Theorem 2. Choose a finite partition of the path space of v_t into sets S_1, \dots, S_n . Choose a representative element v_t^i in each set S_i . Then form an approximate distribution \hat{P} on the path space concentrated on the representative elements, by

$$P[v_t = v_t^i] = P(S_i). \quad (55)$$

The sets S_i should be chosen so that if y_t^i and $y_t^{i'}$ are, respectively, solutions of (53) corresponding to an arbitrary element v_t^i of S_i and the representative element $v_t^{i'}$ of $S_{i'}$, the probability that y_t^i and $y_t^{i'}$ are far apart is small. Then for this finite approximate distribution, use the quantities (51) and (52) of Theorem 3 as approximations for (38) and (39) of Theorem 2.

How, in practice, could the previous results be used to give a control law for the pursuit-evasion problem? First, a finite representative family of control paths that the evader might use or nearly use should be selected and probabilities assigned to them. Then the formulas (51) and (52) of Theorem 3 would give the conditional expectations

$$E[y_t | \mathcal{M}_t] \quad \text{and} \quad E[v_t | \mathcal{M}_t]. \quad (56)$$

Then the conditional expectation of the evader's terminal position is given in terms of the conditional expectations of (56) by (20), the predicted miss by (17), and the control by (7). This gives a computationally feasible procedure for computing (7).

V. CONCLUSIONS

Previous work [1] implied that the optimal control for the pursuit-evasion problem could be given in terms of a formula involving the conditional expectation of the evader's terminal position. The Kallianpur–Striebel formula, a double conditioning with respect to both the past measurements and the values of the evader's unknown control law, and Kalman filtering were used to express this conditional expectation in terms of expected values of given functions of past measurements with respect to the distribution of the unknown control law. When the unknown control law has only a finite number of paths, this gives a computationally feasible procedure for computing the optimal control law. In other cases it gives an approximation to the optimal control law.

REFERENCES

- [1] V. E. Beneš, "Quadratic approximation by linear systems controlled from partial observations," in *Stochastic Analysis. Liber Amicorum of M. Zakai*, M. Merzbach, A. Shwartz, and E. Mayer-Wolf, Eds. New York: Academic, 1992.
- [2] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*. New York: Academic, 1988.
- [3] J. R. Cloutier, J. H. Evers, and J. J. Feeley, "Assessment of Air-to-air missile guidance and control technology," *IEEE Contr. Syst. Mag.*, vol. 9, pp. 27–34, 1989.
- [4] M. H. A. Davis and S. Markus, "An introduction to non-linear filtering," in *Stochastic Systems. The Mathematics of Filtering and Identification*, M. Hazewinkel and J. C. Williams, Eds. Dordrecht, The Netherlands: Reidel, 1981.
- [5] K. L. Helmes and R. W. Rishel, "The solution of a partially observed stochastic control problem in terms of predicted miss," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1462–1464, 1992.

- [6] — An optimal control depending on the conditional density of the unobserved state in *Applied Stochastic Analysis: Lecture Notes in Control and Information Sciences* 177 T. Karatzas and D. Ocone, Eds. New York: Springer Verlag, 1992.
- [7] R. L. Moose et al. Modeling and estimation for tracking maneuvering targets. *IEEE Trans Aerosp Electron Syst*, vol. AES-15, pp. 448–455, May 1979.
- [8] H. L. Pastrick, S. M. Seltzer, and M. E. Warren. Guidance laws for short-range tactical missiles. *J. Guid. Contr.*, vol. 4, no. 2, pp. 98–108, 1981.
- [9] E. Y. Rodin. A pursuit evasion bibliography version 2. *Comput. Math. Appl.*, vol. 18, no. 1–3, pp. 245–320, 1989.
- [10] A. S. Willsky. Detection of abrupt changes in dynamical systems. in *Lecture Notes in Control and Information Sciences*, vol. 77. *Detection of Abrupt Changes in Signals and Dynamical Systems*. M. Basseville and A. Benveniste, Eds. Berlin: Springer Verlag, 1986.

A Subspace Fitting Method for Identification of Linear State-Space Models

A. Swindlehurst, R. Roy, B. Ottersten, and T. Kailath

Abstract—A new method is presented for the identification of systems parameterized by linear state-space models. The method relies on the concept of *subspace fitting*, wherein an input/output data model parameterized by the state matrices is found that best fits, in the least-squares sense, the dominant subspace of the measured data. Some empirical results are included to illustrate the performance advantage of the algorithm compared to standard techniques.

I. INTRODUCTION

Research in the identification of discrete time linear systems has focused in recent years on prediction error methods (PEMs) based on autoregressive and autoregressive moving average (ARMA) data models and their various derivatives (e.g., [1], [2]). Although linear state space models are common in estimation and control, they have not been widely used in identification. Implicitly, of course, a given input/output model can be associated with various canonical state space realizations, and thus a PEM might be thought of as finding a state space model for the system. One of the goals of this note is to demonstrate, however, that there are some important advantages in explicitly considering more general state space forms in identification.

Several early approaches to the general state space identification problem were based on examining the structure of a Hankel matrix composed of samples of the impulse response of the system [3], [5]. More recently, De Moor [6], [7] has developed a total least squares algorithm that exploits the same shift structure present in a certain input/output data model, and that allows arbitrary input excitations.

Manuscript received May 21, 1993; revised February 1, 1994. This work was supported by the National Science Foundation under Grant MIP 9110112.

A. Swindlehurst is with the Department of Electrical and Computer Engineering, Brigham Young University, Provo, UT 84602, USA.

R. Roy is with ArrayComm, Inc., Santa Clara, CA 95054, USA.

B. Ottersten is with the Department of Signals, Sensors, and Systems, Royal Institute of Technology, S-100 44 Stockholm, Sweden.

T. Kailath is with the Information Systems Lab, Stanford University, Stanford, CA 94305, USA.

IEEE Log Number 9406993.

Related methods have also been recently proposed by Moonen [8] and Verhaegen et al. [9].

The method presented in this note also exploits the inherent shift structure in the data, but in a different way. The motivation for this new algorithm comes from some recent results in sensor array signal processing. In particular, it is shown how the identification problem can be cast in the *subspace fitting* framework, where the goal is to find the input/output model which best fits (in the least squares sense) the dominant subspace of the data. This approach has been successfully applied by Ottersten and Viberg in the context of narrowband direction-of-arrival estimation [10]. Of special interest is the fact that a weighting can be applied to the dominant subspace to emphasize certain directions where the signal to noise ratio is high. This weighting can provide an advantage in cases involving nearly unobservable systems or an insufficiently excited state space.

II. DATA MODEL AND ASSUMPTIONS

Consider the following multiple input multiple output (MIMO) time invariant linear system in state space form:

$$x_{k+1} = Ax_k + Bu_k$$

$$y_k = Cx_k + Du_k + v_k \quad (1)$$

where $x_k \in \mathbb{R}$, $u_k \in \mathbb{R}$, $y_k \in \mathbb{R}^l$, $v_k \in \mathbb{R}^l$ and the system matrices A , B , C , and D are of consistent dimension. The system input u_k is assumed known and the output y_k is corrupted by additive measurement noise v_k . If several observations of the input and output vectors are available, they may be grouped together into the single equation (e.g., see [6] and [11])

$$Y = GX + HU + V \quad (2)$$

where

$$Y = \begin{bmatrix} y_k & y_{k+1} & \dots & y_{k+j-1} \\ y_{k+1} & y_{k+2} & \dots & y_{k+j} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k+j-1} & y_{k+j} & \dots & y_{k+j+j-1} \end{bmatrix}$$

$$G = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{j-1} \end{bmatrix}$$

$$H = \begin{bmatrix} D & 0 & 0 & 0 \\ CB & D & 0 & 0 \\ CAB & CB & D & 0 \\ CA^2B & CA^3B & CA^4B & D \end{bmatrix}$$

$x = [x_k x_{k+1} \dots x_{k+j-1}]$ and where $U(m \times j)$ and $V(l \times j)$ are block Hankel matrices constructed exactly as Y but containing samples of the input and disturbance sequences, respectively.

We will make the following assumptions concerning the data model of (2):

- The system is observable, and the block dimension l is chosen to be large enough so that $\text{rank}(G) = n$. This implies $li \geq n$.
- The input sequence u_k has sufficiently excited the system and $j \geq n$, so that $\text{rank}(X) = n$.

- The input is uncorrelated with the measurement noise (open-loop operation).
- A nontrivial matrix U^\perp can be found to satisfy $UU^\perp = 0$. This requires $j > mi$, and leads to

$$YU^\perp = \Gamma XU^\perp + VU^\perp. \quad (3)$$

- $\text{rank}(XU^\perp) = n$, so that $j \geq mi + n$. This rank condition is satisfied for most choices of the input (e.g., with probability one if the input is zero-mean white noise) [6], [11].

With these assumptions, YU^\perp is an $li \times p$ matrix, with $n \leq p \leq j - mi$. In most situations, $p = j - mi > n$, so that YU^\perp is composed of a (low) rank n "signal" term ΓXU^\perp that depends on the parameters of interest, and a "noise" term VU^\perp that is (with probability one) full rank.

It is convenient for us to also define here the "sample covariance" matrix

$$\hat{R}_{YU^\perp} = \frac{1}{j} (YU^\perp)(YU^\perp)^T$$

and note that it converges to its limiting expected value with probability one

$$R_{YU^\perp} = \lim_{j \rightarrow \infty} \mathcal{E} \{ \hat{R}_{YU^\perp} \} = \Gamma R_{XU^\perp} \Gamma^T + \sigma^2 \Sigma$$

where

$$R_{XU^\perp} = \frac{1}{j} (XU^\perp)(XU^\perp)^T, \quad \sigma^2 \Sigma = \frac{1}{j} \mathcal{E} \{ (VU^\perp)(VU^\perp)^T \}$$

and Σ is normalized so that $\det(\Sigma) = 1$. It is easy to show that the generalized eigenvalues $\{\lambda_k\}$ of the pair (R_{YU^\perp}, Σ) satisfy $\lambda_1 \geq \dots \geq \lambda_n > \lambda_{n+1} = \dots = \lambda_l = \sigma^2$. Furthermore, the eigenvectors $E = [e_1 \dots e_n]$ associated with $\lambda_1, \dots, \lambda_n$ satisfy

$$\Sigma E = \Gamma(\eta)T \quad (4)$$

for some $n \times n$ matrix T . We have written Γ as a function of a vector η , which contains the elements of the parameterization chosen for A and C . Just as there are many realizations or coordinate systems that can be used to describe the state space, there are many identifiable parameterizations η that can be chosen, each yielding a different T that satisfies (4). The subspace fitting method presented in this note is based on the relationship of (4).

Note that with no measurement noise or infinite data, the model order n is revealed by simply counting how many of the smallest generalized eigenvalues are equal. With a finite collection of noisy data, a statistical test is required to estimate this quantity. This problem has been extensively studied in very general contexts, and many such tests have been developed (e.g., see [12]–[14]). We will thus assume throughout the remainder of the note that n has been correctly determined.

III. A SUBSPACE FITTING APPROACH

Because of the effects of noise, only an estimate of the generalized eigenvectors \hat{E} can be obtained from \hat{R}_{YU^\perp} (or equivalently from an SVD of YU^\perp). Consequently, assuming an appropriate identifiable parameterization η has been chosen, we propose the following two-step identification procedure based on (4).

1) Estimate η by means of the following weighted least-squares problem:

$$\begin{aligned} \hat{\eta} &= \arg \min_{\eta} \| \Sigma \hat{E} W^{1/2} - \Gamma(\eta)T \|^2_F \\ &= \arg \min_{\eta} \text{Tr}(P_{\hat{\eta}}^\perp \Sigma \hat{E} W \hat{E}^T \Sigma^T) \end{aligned} \quad (5)$$

where $P_{\hat{\eta}}^\perp = I - \Gamma(\hat{\eta})[\Gamma(\hat{\eta})^T \Gamma(\hat{\eta})]^{-1} \Gamma(\hat{\eta})^T$. The estimates of the system matrices are thus given by $A(\hat{\eta}), C(\hat{\eta})$.

2) Estimate B and C using a noise-free version of (2) and the estimates $A(\hat{\eta}), C(\hat{\eta})$.

Some comments are in order concerning the two steps of the Weighted Subspace Fitting (WSF) algorithm described above.

Step 1: The choice of the weighting matrix W will be discussed below in Subsection A. A variety of possible parameterizations for η exist, depending on what (if any) prior information about the system is available. For example, if the system is multiple-input single-output, a particularly appropriate choice is to assume that the system is in observer form, and hence that A is left-companion and $C = [1 \ 0 \dots 0]$. Alternatively, A could be assumed to be diagonalizable, in which case C is solved for directly, and the minimization of (5) can be simplified to involve only a search over the n pole locations.

One of the advantages of the problem formulation considered herein is that any *a priori* information about the structure of A and C can be directly incorporated into the problem. In some instances, the dynamical model underlying the system may be well understood, but may depend on certain parameters that are imprecisely known. These parameters appear as unknown constants in the matrices A and C of the system model, and could serve as the parameters in the minimization of (5) just as easily as the poles or the coefficients of the characteristic polynomial. Whether or not the set of unknown parameters can be uniquely identified is then problem dependent, and must be determined on a case-by-case basis.

The minimization of (5) is identical in form to a problem in antenna array signal processing considered in [15]. In fact, once an appropriate parameter vector has been identified, an algorithm essentially identical to that described in [15] may be used to estimate A and C . Additional information on the connection between state-space system identification and array signal processing can be found in [16] and [17].

We note here that, even though $\Sigma \neq I$, consistent parameter estimates can be obtained by setting $\Sigma = I$ if the noise is white and independent with equal variance among the l outputs, and if j is sufficiently large.¹ In other cases, however, ignoring Σ when $\Sigma \neq I$ (as is done in [6], [7], [9]) can lead to biased parameter estimates.

Step 2: This step is also used by De Moor in his identification method [6], [7]. A description of the mechanics required to solve for B and D can also be found in [19], and will not be given here.

A. Subspace Weighting

In simple terms, the presence of the weighting matrix W in the minimization of (5) allows certain directions or dimensions of the low rank subspace to be emphasized over others. For example, the generalized eigenvector e_1 corresponding to the largest eigenvalue represents the component of the measured data where the signal energy is the strongest. The eigenvector e_n gives the direction where the signal energy is weakest but still nonzero. A measure of how "strong" or "weak" the signal energy is in a given direction can be based on how much bigger than σ^2 the eigenvalues $\lambda_1, \dots, \lambda_n$ are. The difference $\lambda_n - \sigma^2$ can be quite small in cases where Γ or X are ill-conditioned (e.g., because the system is nearly unobservable or the input is not sufficiently exciting), and in such cases one would probably wish to give much less weight to e_n in (5) than e_1 .

The following "optimal" weighting has been derived for subspace fitting problems of the form (5) in [10]:

$$W_{\text{WSF}} = (\hat{\Lambda}_n - \sigma^2 I)^2 \hat{\Lambda}_n^{-1} \quad (6)$$

¹How large j has to be for this to hold depends on the magnitude of the noise.

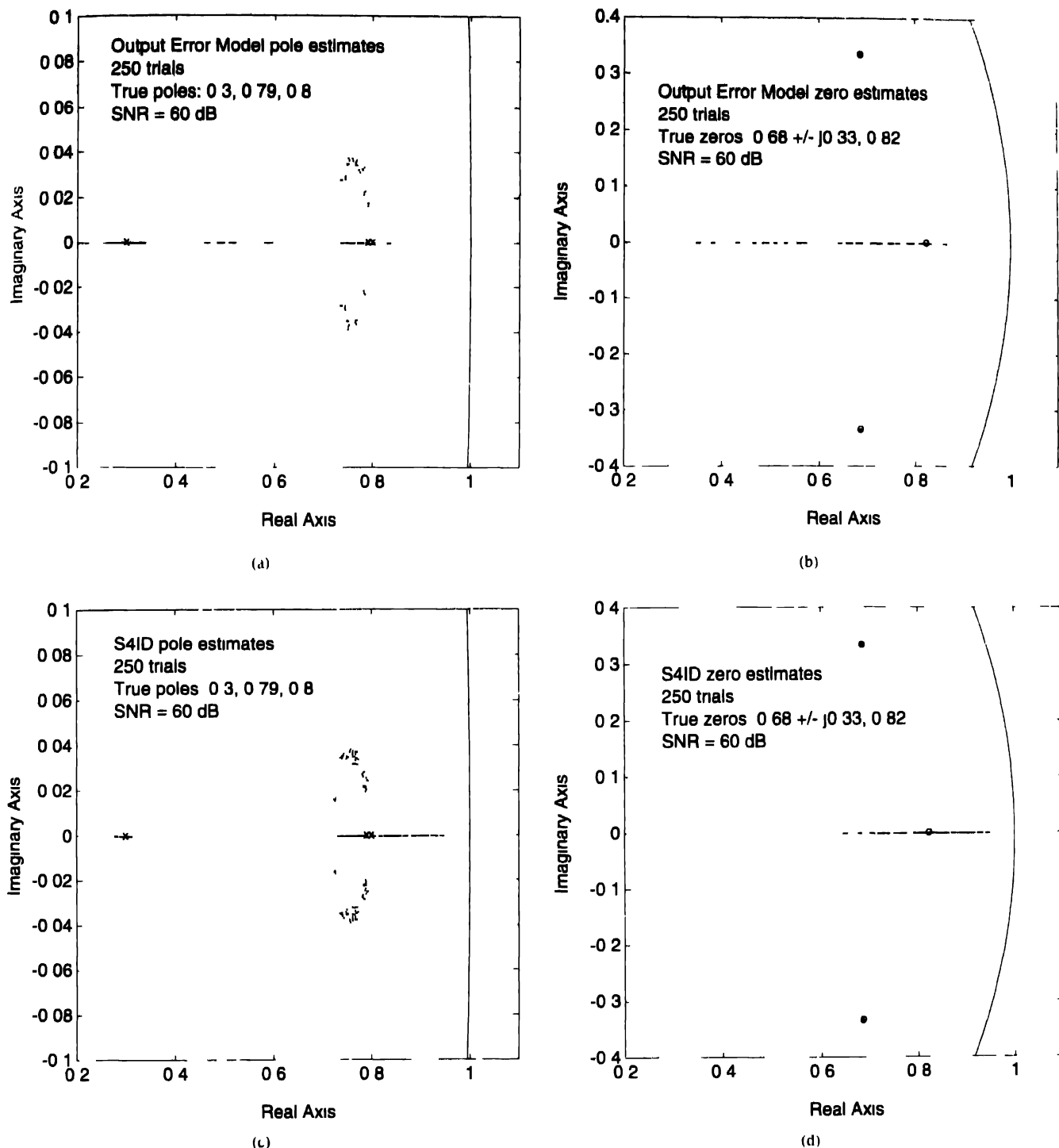


Fig 1 Pole and zero estimates for simulation example 1

where the diagonal matrix \mathbf{A} contains $\lambda_1, \dots, \lambda_n$ as its diagonal elements, and σ^2 is an appropriate estimate of σ^2 (obtained, for example as an average of the $l-n$ smallest generalized eigenvalues). The term "optimal" here means that, under certain conditions, the weighting of (6) will lead to minimum variance parameter estimates. The Hankel structure of \mathbf{Y} and \mathbf{V} in our problem violates one of these conditions, namely, that the columns of the data matrix be independent. However, using \mathbf{W}_{WSF} in (5) has been empirically shown to yield better results than using no weighting at all ($\mathbf{W} = \mathbf{I}$).

B Comparison to Other Techniques

The principle difference between the WSF algorithm presented above and the methods described in [6], [7], and [9] is that the WSF approach can exploit the complete structure of \mathbf{F} rather than just a single "shift-invariance." This advantage is especially evident in situations where *a priori* information is available about the structure of \mathbf{A} and \mathbf{C} . On the other hand, the methods of [6], [7], and [9] admit a simple "closed-form" solution, whereas (5) can only be solved by means of a multidimensional search.

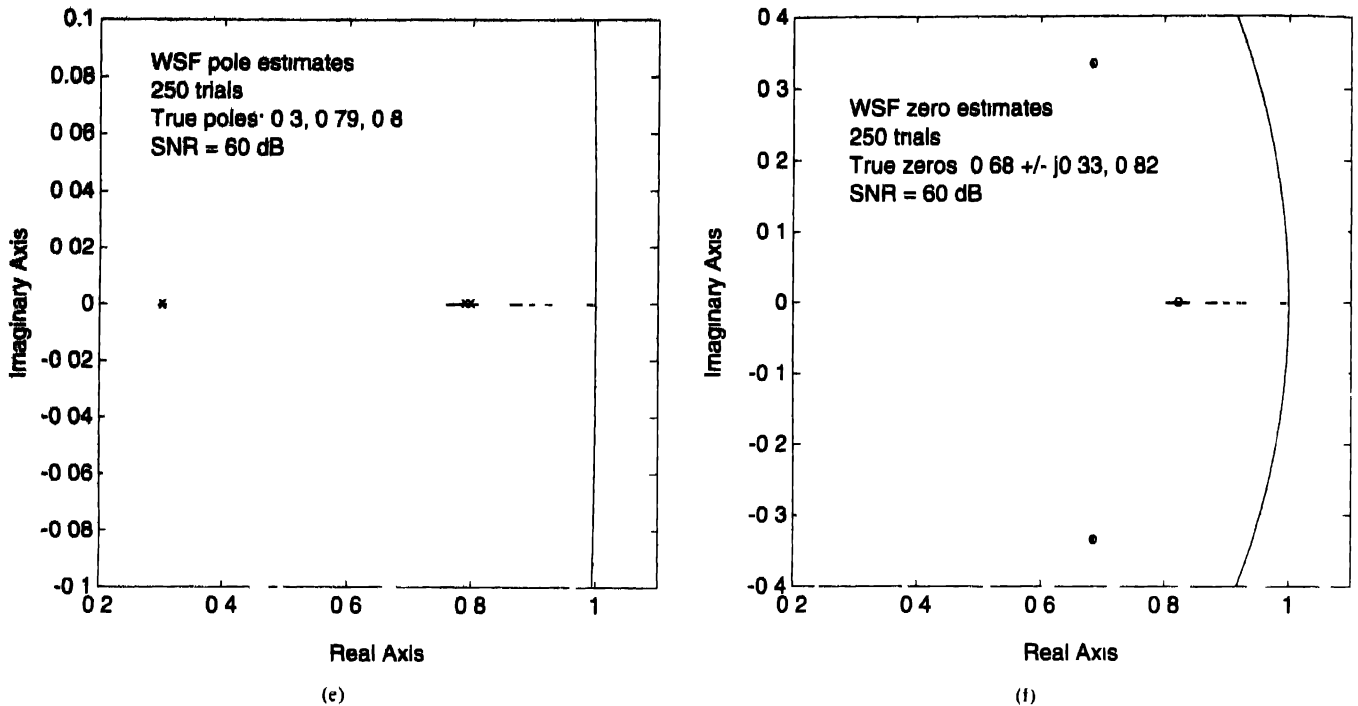


Fig. 1 Continued

Unlike standard ARMA-based PEM approaches, the WSF algorithm is just as easily applied to MIMO systems as in the single-input single-output case, has a built-in mechanism for model order determination, and can exploit structured system models with more compact parameterizations. Even in the unstructured case, WSF tends to yield pole and zero estimates with much lower variance than PEM's, as illustrated by the simulation examples of the next section. Recent work in identification and controller design for rapid thermal processing of semiconductor wafers [19], [20] has also borne out some of the advantages of using subspace-based identification methods over PEM's.

IV. SIMULATION EXAMPLES

In this section we consider two simple simulation examples that illustrate the advantage of the subspace fitting approach. In the first example, the parameters of the following three state SISO discrete time system will be estimated:

$$A = \begin{bmatrix} 0.30 & -0.40 & 0.60 \\ 0.00 & 0.80 & 0.00 \\ 0.00 & 0.00 & 0.79 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

$$C = [1.30 \quad 0.40 \quad -0.80]$$

$$D = 1$$

The poles of this system are located at 0.3, 0.79, and 0.8, and the zeros at 0.82, and $0.68 \pm j0.33$. Note that for this system, X will be ill-conditioned since the lower right 2×2 block of A is diagonal, with nearly equal diagonal elements, and the last two elements of B are identical. This is also evidenced by the fact that there is a near pole-zero cancellation in the transfer function.

This system was simulated using a zero mean unit power white Gaussian process as its input, and assuming a zero initial state.² White Gaussian measurement noise with a standard deviation of 0.001 was added to the system output. Using the noise free input and the noisy system output, three methods were used to estimate the system poles and zeros. These were the method of De Moor [6], [7] (which, for brevity, we refer to as the S4ID technique), the WSF algorithm, and a PEM based on an *output error* model [1]. The correct model order was assumed to be known in each case.

We conducted 250 Monte Carlo experiments with an independent measurement noise and input sequence generated for each trial. The block dimensions of the Hankel matrices were chosen as $i = 12$ and $j = 50$, corresponding to a total of 61 output samples for each trial. The WSF method was implemented using the weighting of (6) and assuming a diagonal parameterization for the system matrix A . Both WSF and PEM used the true parameter vectors to initialize their respective search routines. The results of the simulations are displayed in Fig. 1. The solid line at the right of each plot represents the unit circle, and the true pole and zero locations are indicated by the symbols \times and \circ , respectively.

The variance of the poles and zeros estimated by WSF is clearly much smaller than that of the other algorithms. This is especially true for the pole at 0.3, in fact, all 250 estimates are so closely bunched together that they are almost indistinguishable from the \times marking the true pole. All three algorithms estimated the complex zeros very accurately, the individual trials are again indistinguishable from the true zero locations. However, the variance of the real valued zero is much smaller for WSF than for either PEM or S4ID. In addition, both PEM and S4ID estimated complex poles on about 20% of the simulations, while the WSF poles were always purely real (although they were not constrained to be so). One drawback of the subspace based methods is that, on a few occasions (5 for WSF and 11 for S4ID), they produced unstable pole estimates.

²The effects of a nonzero initial state can be handled by assuming the presence of an additional input that is a unit Dirac impulse [6].

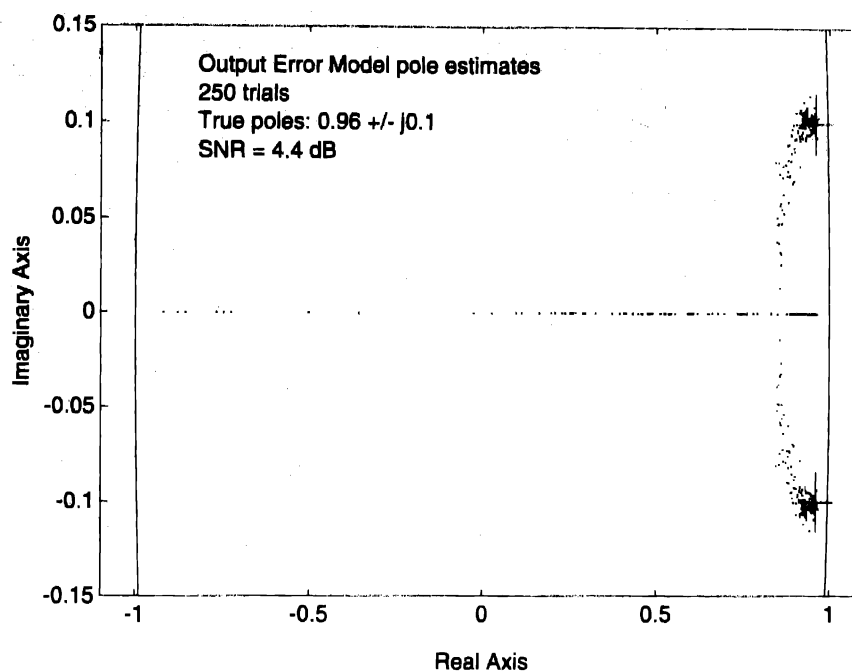


Fig. 2. Pole estimates for output error model, example 2.

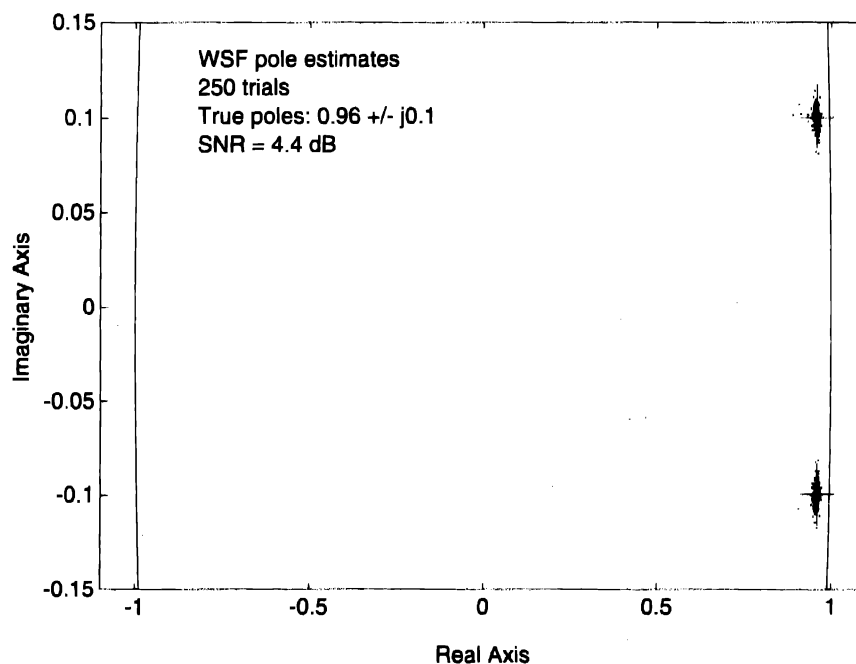


Fig. 3. Pole estimates for WSF, example 2.

In the second example, we consider the following system used in the simulation studies of [9]

$$A = \begin{bmatrix} 1.92 & -0.9316 \\ 1 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$C = [0.05 \quad 0.025]$$

$$D = 0.$$

The poles of this system are at $0.96 \pm j0.01$, and there is a single zero at -0.5 . The same type of input and measurement disturbance as above were used in this case, except the variance of the disturbance was increased to one. This resulted in an average effective output SNR of only 4.4 dB. The dimensions of the block Hankel matrices were chosen as $i = 40$ and $j = 161$, for a total of 200 samples per trial. As before, both WSF and the output error PEM were initialized with the true parameters, and 250 independent trials were conducted. The results of the simulation are plotted in Figs. 2 and 3.

Fig. 2 shows the pole estimates for the output error model, with the true pole locations indicated by the crosshairs, and Fig. 3 shows

the same for the WSF algorithm. No separate plot is shown for S4ID, since it gave results essentially equivalent to WSF in this case (with the exception of one outlier). The WSF approach again yields pole estimates with a much smaller variance than the corresponding PEM. The average mean-square prediction error for WSF in this case was 0.985, compared to 1.403 for the output error PEM.

Because of the various optimality properties of PEM's in general [1], one may be surprised by the superior performance of WSF in the above two examples. Since both algorithms used essentially the same minimization procedure (Gauss-Newton iterations with identical initial conditions and termination criteria) the relatively poor performance of the output error PEM in these examples may be attributed to a propensity for convergence to local minima.

V CONCLUSIONS

We have presented a new method for identification of linear systems parameterized by state-space models. The method is based on the notion of weighted subspace-fitting (WSF), a problem-solving philosophy most recently applied to parameter estimation problems in sensor array signal processing. Like other state space identification schemes, WSF is easily applied in the MIMO case, allows for direct incorporation of *a priori* physical constraints into the problem, and appears to have better numerical properties than methods based on input/output models. In addition, the ability to appropriately weight the signal subspace vectors used in the WSF minimization provides a degree of robustness over previous subspace-based methods when the system is nearly unobservable or not sufficiently excited. Although implementation of WSF requires a multidimensional parameter search, accurate initial conditions for the search are readily obtained.

REFERENCES

- [1] L. Ljung, *System Identification—Theory for the User*, Englewood Cliffs, NJ: Prentice Hall, 1987.
- [2] T. Söderström and P. Stoica, *System Identification*, Englewood Cliffs, NJ: Prentice Hall, 1989.
- [3] B. L. Ho and R. E. Kalman, "Efficient construction of linear state variable models from input/output functions," *Regelungstechnik*, vol. 14, pp. 545–548, 1966.
- [4] H. P. Zeiger and A. J. McEwen, "Approximate linear realization of given dimension via Ho's algorithm," *IEEE Trans Automat Contr*, vol. AC-19, p. 153, 1974.
- [5] S. Y. Kung, "A new identification and model reduction algorithm via singular value decomposition," in *Proc. 12th Asilomar Conf. Circuits Syst. Comp.*, Asilomar, CA, Nov. 1978, pp. 705–714.
- [6] B. De Moor, "Mathematical concepts and techniques for modelling of static and dynamic systems," Ph.D. dissertation, Katholieke Universiteit Leuven, Leuven, Belgium, 1988.
- [7] B. De Moor, M. Moonen, L. Vandenbergh, and J. Vandewalle, "A geometrical approach for the identification of state space models with singular value decomposition," in *Proc. IEEE ICASSP*, vol. 4, New York, 1988, pp. 2244–2247.
- [8] M. Moonen, B. De Moor, L. Vandenbergh, and J. Vandewalle, "On and off line identification of linear state space models," *Int. J. Contr.*, vol. 49, no. 1, pp. 219–232, 1989.
- [9] M. Verhaegen and P. DeWilde, "Subspace model identification, Part 1–2," *Int. J. Contr.*, vol. 56, no. 5, pp. 1187–1241, 1992.
- [10] M. Viberg and B. Ottersten, "Sensor array processing based on subspace fitting," *IEEE Trans. Signal Processing*, vol. 39, pp. 1110–1121, May 1991.
- [11] B. Gopinath, "On the identification of linear time invariant systems from input-output data," *Bell Syst. Tech. J.*, vol. 48, pp. 1101–1113, 1969.
- [12] T. W. Anderson, "Asymptotic theory for principal component analysis," *Ann. Math. Statist.*, vol. 34, pp. 122–148, 1963.
- [13] H. Akaike, "A new look at statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.
- [14] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-33, pp. 387–392, Apr. 1985.
- [15] A. Swindlehurst, B. Ottersten, R. Roy, and T. Kailath, "Multiple invariance ESPRIT," *IEEE Trans. Signal Processing*, vol. 40, pp. 867–881, Apr. 1992.
- [16] A. Swindlehurst, R. Roy, B. Ottersten, and T. Kailath, "System identification via weighted subspace fitting," in *Proc. Amer. Contr. Conf.*, June 1992, pp. 2158–2163.
- [17] A. van der Veen, F. Deprittere, and A. Swindlehurst, "Subspace based signal analysis using singular value decomposition," *Proc. IEEE*, vol. 81, pp. 1277–1308, Sept. 1993.
- [18] M. Viberg, B. Ottersten, B. Wahlberg, and L. Ljung, "A statistical perspective on state space modeling using subspace methods," in *Proc. 30th CDC Conf.*, Brighton, England, 1991.
- [19] P. Gyugyi, Y. M. Cho, G. Franklin, T. Kailath, and R. Roy, "A model based control of rapid thermal processing systems," in *Proc. 1st IEEE Conf. Contr. Appl.*, Dayton, OH, Sept. 1992.
- [20] P. Gyugyi, Y. M. Cho, G. Franklin, and T. Kailath, "Control of water temperature in rapid thermal processing. Part I—State space model identification and control," submitted to *Automatica*, 1992.

Consistency of Modified LS Estimation Method for Identifying 2-D Noncausal SAR Model Parameters

Ping Ya Zhao and John Litva

Abstract—Least squares (LS) and maximum likelihood (ML) are the two main methods for parameter estimation of two-dimensional (2-D) noncausal simultaneous autoregressive (SAR) models. ML is asymptotically consistent and unbiased but computationally unattractive. On the other hand, conventional LS is computationally efficient but does not produce accurate parameter estimates for noncausal models. Recently, in [17], a modified LS estimation method was proposed and shown to be unbiased. In this note we prove that, under certain assumptions, the method introduced in [17] is also consistent.

I INTRODUCTION

Two dimensional (2-D) signal and system modeling and parameter estimation have many applications such as 2-D Kalman filtering [1], image estimation and identification [2]–[4], image restoration [5], [6], multidimensional spectrum estimation [7], [8], direction finding [9], texture analysis and synthesis [10], [11], multidimensional system identification [12], [13], etc. Several kinds of models are used to cite a few: Gauss-Markov random field (GMRF) models which are sometimes called conditional Markov (CM) models [14]–[16], simultaneous autoregressive (SAR) models [16], [17], autoregressive moving average (ARMA) models [18], [19], etc. In this note we will concentrate on the problem of estimating the parameters $\{q_k\}$ of the 2-D noncausal SAR model in the form of

$$y(i, j) = - \sum_{(k, l) \in N} q_k y(i-k, j-l) + e(i, j) \quad (1)$$

Manuscript received June 18, 1993; revised February 20, 1994. The authors are with Communications Research Laboratory, McMaster University, Hamilton, Ontario, Canada L8S 4K1. IEEE Log Number 9406994.

where $\{\epsilon(t, j)\}$ is a field of independent and identically distributed random variables with zero mean and constant variance. Λ is called the neighbor set of the model whose shape determines the causality of the model. According to [20] the model is a noncausal one if Λ is the form

$$\Lambda \subseteq \{(k, l) : -p \leq k \leq p, -q \leq l \leq q\} \setminus \{(0, 0)\} \quad (2)$$

This is the most general of all the models and is known to provide the best performance. For the same specific problem using the noncausal model will lead to simpler model form and more efficient computational schemes.

If the model is noncausal however $\epsilon(t, j)$ will correlate with $\{q(t-k, j-l) : (k, l) \in \Lambda\}$ and the conventional least squares (LS) will not be an unbiased and consistent parameter estimator. One is then forced to use the computationally inefficient maximum likelihood (ML) estimator to determine the noncausal SAR model parameters accurately. To avoid excessive computation of the ML an iterative scheme was developed in [16] which gives approximate ML estimates of the SAR model parameters. As a means of overcoming inaccuracy in the conventional LS a modified LS estimator was proposed in [17]. The modified LS has been proved mathematically to be an unbiased estimator under some mild conditions and the experiments in [17] show that the modified LS and the approximate ML have the same estimation accuracy. On the other hand the modified LS is superior to the approximate ML from a computational point of view. If the modified LS estimate is calculated directly from (14) it requires fewer operations than for one iteration of the approximate ML whose implementation usually requires several iterations. More importantly the modified LS possesses some shift invariance properties [21] which can be used to develop a fast implementation based on recursion. An order recursive algorithm with the computational complexity of $O(m^2)$ where m is the number of the estimated parameters is given in [17] and a spatially recursive algorithm with the computational complexity of $16m^{1/2} + O(m)$ multiplications and divisions per recursion (MADPR) is given in [22]. The spatially recursive algorithm uses only the local observations appearing in a sliding data window so it can be used to estimate 2-D nonstationary and noncausal SAR model parameters. It is believed that the computational efficiency is an important issue and recursive algorithms usually lead to reduction in computational burden [23] [24].

When further comparisons are made to the conventional LS estimator the modified LS estimator is favored to have another significant advantage. The conventional LS normal equations are sometimes ill conditioned which leads to numerical instabilities. However the coefficient matrix of the modified LS normal equation has a different structure with a lower probability of being singular. It thus can effectively avoid the instability problem which is very important in real implementations of the parameter estimators.

Yet consistency is an extremely essential characteristic for an estimator which can serve as a very important criterion for selecting an estimator for a particular application. The mathematical proof of the consistency of the above mentioned modified LS estimator has not been presented anywhere in the literature. The purpose of this note is to present a proof for the consistency of the unbiased modified LS estimator. This proof corroborates computer simulation results that the mean square errors of the modified LS decrease at their greatest rate when the sizes of the data sample windows increase and that the mean square error of the modified LS estimator has a rate of decrease which is much greater than the approximate ML.

II THE MODIFIED LS ESTIMATION METHOD

When the model is a causal one the conventional LS is an unbiased and consistent estimator. The basic idea of the modified LS is the translation of the noncausal SAR model parameter estimation problem into a problem based on a causal model and an anticausal model. This is achieved by translating (1) into

$$q_{0,0}y(t, j) = - \sum_{(k,l) \in \Lambda} q_{k,l}y(t-k, j-l) + \epsilon(t, j) \quad (3)$$

$$q_{0,0}y(t, j) = - \sum_{(k,l) \in \Lambda} q_{k,l}y(t-k, j-l) + \epsilon(t, j) \quad (4)$$

where

$$q_{0,1} + q_{0,-1} - q_{0,0} = 1 \quad (5)$$

$$\epsilon(t, j) + \epsilon(t, j) = \epsilon(t, j) \quad (6)$$

$$\Lambda \subseteq \{(k, l) : 1 \leq k \leq p, -q < l \leq q\} \cup \{(0, l) : 1 < l \leq q\} \quad (7)$$

$$\Lambda \subseteq \{(k, l) : -p \leq k < -1, -q < l \leq q\} \cup \{(0, l) : -q < l \leq -1\} \quad (8)$$

provided that such a decomposition exists. Minimization of the modified LS objective function

$$\begin{aligned} F(\mathbf{g}_i) &= \sum_{(t,j) \in G} \{[\epsilon(t, j)]^2 + [\epsilon(t, j)]^2\} + 2D(q_{0,1} + q_{0,-1} - 1) \\ &= \sum_{(t,j) \in G} \{[\epsilon(t, j)]^2 + [\epsilon(t, j)]^2\} + 2n_G d(q_{0,1} + q_{0,-1} - 1) \end{aligned} \quad (9)$$

with respect to \mathbf{g}_i results in the modified LS estimator where G is a 2-D data window, n_G is a number of data points contained in G and

$$\mathbf{g}_i = [\mathbf{g}_i^T \text{ ml } d \mathbf{g}_i^T]^T \quad (10)$$

$$\mathbf{g}_i = \left[\text{column} \{q_{k,l} : (k,l) \in \Lambda\} \right] \quad (11)$$

$$\mathbf{g}_i = \left[\text{column} \{q_{k,l} : (k,l) \in \Lambda\} \right] \quad (12)$$

$$D = n_G d = - \frac{\sum_{(t,j) \in G} [\epsilon(t, j)]}{q_{0,0}} = - \frac{\sum_{(t,j) \in G} [\epsilon(t, j)]}{I_0} \quad (13)$$

The resulting modified LS estimator is

$$\mathbf{g}_i(G) = \mathbf{R}_i^{-1}(G) \mathbf{z}_i \quad (14)$$

where

$$\mathbf{R}_i(G) = \begin{bmatrix} \mathbf{J} \mathbf{R}_i(G) \mathbf{J} & \mathbf{0} & \mathbf{0} \\ & n_G & \\ \mathbf{0} & n_G & \mathbf{0} \\ & n_G & \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_i(G) \end{bmatrix} \quad (15)$$

$$\mathbf{R}_p^c(G) = \sum_{(i,j) \in G} \mathbf{x}_p^c(i,j) \mathbf{x}_p^{cT}(i,j) \quad (16)$$

$$\mathbf{R}_p^a(G) = \sum_{(i,j) \in G} \mathbf{x}_p^a(i,j) \mathbf{x}_p^{aT}(i,j) \quad (17)$$

$$\mathbf{x}_p^c(i,j) = \begin{bmatrix} y(i,j) \\ \mathbf{y}_p^c(i,j) \end{bmatrix} \quad (18)$$

$$\mathbf{x}_p^a(i,j) = \begin{bmatrix} y(i,j) \\ \mathbf{y}_p^a(i,j) \end{bmatrix} \quad (19)$$

$$\mathbf{y}_p^c(i,j) = \text{column}\{y(i-k, j-l) : (k,l) \in N^c\} \quad (20)$$

$$\mathbf{y}_p^a(i,j) = \text{column}\{y(i-k, j-l) : (k,l) \in N^a\} \quad (21)$$

$$\mathbf{x}(i,j) = [y(i,j+q) \cdots y(i,j) \cdots y(i,j-q)]^T \quad (22)$$

$$\mathbf{z}_p = [0 \cdots 0 n_G 0 \cdots 0]^T. \quad (23)$$

Boldface $\mathbf{0}$ stands for a matrix or vector with all zero entries, \mathbf{J} is a variable dimension square exchange matrix with ones along the antidiagonal and zeros everywhere else, and $\mathbf{g}_p(G)$ denotes the estimate of \mathbf{g}_p based on the use of the data $\{y(i,j) : (i,j) \in B\}$, and $B \supset G$ is another data window, such that for $(i_1, j_1) \in G$, the elements of vectors $\mathbf{x}_p^c(i_1, j_1)$ and $\mathbf{x}_p^a(i_1, j_1)$ are defined on B , i.e., if $y(i,j)$ is an element of $\mathbf{x}_p^c(i_1, j_1)$ or $\mathbf{x}_p^a(i_1, j_1)$, then $(i,j) \in B$.

Using the same procedure as that in [17], we can prove that $\mathbf{g}_p(G)$ is an unbiased estimate of \mathbf{g}_p if

- 1) $\mathbf{R}_p(G)$, defined by (15), is a positive matrix and independent of $\mathbf{g}_p - \mathbf{g}_p(G)$;
- 2) the decomposition of (1) into (3) and (4) and $e(i,j)$ into $e^c(i,j)$ and $e^a(i,j)$ which are independent and have zero means and constant variances exists, and (13) is valid for sufficiently large G ; and
- 3) $e^c(i,j)$ and $e^a(i,j)$ are uncorrelated with $\{y(i-k, j-l) : (k,l) \in N^c\}$ and $\{y(i-k, j-l) : (k,l) \in N^a\}$, respectively.

Because the same modified LS objective function (9) is used, $\mathbf{g}_p(G)$ is actually the same estimate as that in [17] except that D is factored into $n_G d$ and d , instead of D , is used as the central element of $\mathbf{g}_p(G)$. From (13), we see that d is approximately a constant and D is proportional to n_G if $\{e^c(i,j)\}$ and $\{e^a(i,j)\}$ are stationary.

III. CONSISTENCY OF THE MODIFIED LS ESTIMATOR

Lemma: Given a stationary image $\{y(i,j) : (i,j) \in B\}$, the absolute value of any element of $E\{\mathbf{R}_p(G)[\mathbf{g}_p - \mathbf{g}_p(G)][\mathbf{g}_p - \mathbf{g}_p(G)]^T \mathbf{R}_p^T(G)\}$ does not exceed $O(n_G)$ as $n_G \rightarrow \infty$, where n_G is a number of data points contained in G , if the conditions 2) and 3) are valid, and

- 4) there is a positive number M such that for (i_1, j_1) , (i_2, j_2) , $(i_3, j_3) \in B$ and $k_1 + k_2 + k_3 = 4$,

$$|E\{[y(i_1, j_1)]^{k_1} [e^c(i_2, j_2)]^{k_2} [e^a(i_3, j_3)]^{k_3}\}| \leq M; \quad (24)$$

and

- 5) $e^c(i,j)$ and $e^a(i,j)$ are uncorrelated with $y(i-k, j-l)$ for $(k,l) \notin N \cup \{(0,0)\}$, where N is the neighbor set used in (1).

Proof: By the same procedure as that in [17], it can be shown that

$$\mathbf{R}_p(G)[\mathbf{g}_p - \mathbf{g}_p(G)] = \begin{bmatrix} \sum_{(i,j) \in G} \mathbf{J} \mathbf{y}_p^a(i,j) e^a(i,j) \\ \sum_{(i,j) \in G} y(i,j) e^a(i,j) + D \\ 0 \\ \sum_{(i,j) \in G} y(i,j) e^c(i,j) + D \\ \sum_{(i,j) \in G} \mathbf{y}_p^c(i,j) e^c(i,j) \end{bmatrix}. \quad (25)$$

Let

$$\mathbf{R}_p(G)[\mathbf{g}_p - \mathbf{g}_p(G)][\mathbf{g}_p - \mathbf{g}_p(G)]^T \mathbf{R}_p^T(G) = \mathbf{S} \quad (26)$$

where

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \mathbf{0} & \mathbf{S}_{14} & \mathbf{S}_{15} \\ \mathbf{S}_{12}^T & \mathbf{S}_{22} & \mathbf{0} & \mathbf{S}_{24} & \mathbf{S}_{25} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{S}_{14}^T & \mathbf{S}_{24} & \mathbf{0} & \mathbf{S}_{44} & \mathbf{S}_{45} \\ \mathbf{S}_{15}^T & \mathbf{S}_{25} & \mathbf{0} & \mathbf{S}_{45}^T & \mathbf{S}_{55} \end{bmatrix}. \quad (27)$$

The purpose of the proof is to show that the absolute values of all elements of matrices $E\{\mathbf{S}_{11}\}$, $E\{\mathbf{S}_{15}\}$, and $E\{\mathbf{S}_{55}\}$, vectors $E\{\mathbf{S}_{12}\}$, $E\{\mathbf{S}_{14}\}$, $E\{\mathbf{S}_{25}\}$, and $E\{\mathbf{S}_{45}\}$, as well as scales $E\{\mathbf{S}_{22}\}$, $E\{\mathbf{S}_{24}\}$, and $E\{\mathbf{S}_{44}\}$, do not exceed $O(n_G)$ as $n_G \rightarrow \infty$. Let us begin by studying $E\{\mathbf{S}_{11}\}$.

From condition 2), we conclude that $e^a(i,j)$ has zero mean, and from condition 3) that $e^a(i,j)$ is uncorrelated with $\mathbf{y}_p^a(i,j)$. Further, from condition 5), we conclude that $e^a(i,j)$ is uncorrelated with $y(i-k, j-l)$ for $(k,l) \notin N \cup \{(0,0)\}$. It follows, by using condition 4) for (k_1, l_1) , $(k_2, l_2) \in N^a$ that

$$|E\{y(i_1 - k_1, j_1 - l_1) e^a(i_1, j_1) y(i_2 - k_2, j_2 - l_2) e^a(i_2, j_2)\}| \leq M \quad (28)$$

if $(i_1, j_1) = (i_2, j_2)$, $(i_1 - i_2 + k_2, j_1 - j_2 + l_2) \in N^c \cup \{(0,0)\}$, or $(i_2 - i_1 + k_1, j_2 - j_1 + l_1) \in N^c \cup \{(0,0)\}$; otherwise, the left-hand side of (28) equals zero. Thus, the absolute value of the $(m_N - |k_1(2q+1) + l_1| + 1, m_N - |k_2(2q+1) + l_2| + 1)$ th element of $E\{\mathbf{S}_{11}\}$ can be expressed in the following form:

$$\begin{aligned} & \left| E \left\{ \sum_{(i_1, j_1) \in G} y(i_1 - k_1, j_1 - l_1) e^a(i_1, j_1) \right. \right. \\ & \quad \cdot \left. \sum_{(i_2, j_2) \in G} y(i_2 - k_2, j_2 - l_2) e^a(i_2, j_2) \right\} \Big| \\ & \leq \sum_{\substack{(i_1, j_1), (i_2, j_2) \in G \\ (i_1 - i_2 + k_2, j_1 - j_2 + l_2) \in N^c \cup \{(0,0)\}}} |E\{y(i_1 - k_1, j_1 - l_1) \\ & \quad \cdot e^a(i_1, j_1) y(i_2 - k_2, j_2 - l_2) e^a(i_2, j_2)\}| \\ & + \sum_{\substack{(i_1, j_1), (i_2, j_2) \in G \\ (i_2 - i_1 + k_1, j_2 - j_1 + l_1) \in N^c \cup \{(0,0)\}}} |E\{y(i_1 - k_1, j_1 - l_1) \\ & \quad \cdot e^a(i_1, j_1) y(i_2 - k_2, j_2 - l_2) e^a(i_2, j_2)\}| \\ & + \sum_{(i_1, j_1) \in G} |E\{y(i_1 - k_1, j_1 - l_1) \\ & \quad \cdot y(i_1 - k_2, j_1 - l_2) [e^a(i_1, j_1)]^2\}| \\ & = S \leq 2(m_N + 1)n_G M + n_G M \end{aligned} \quad (29)$$

where m_N is the number of the parameters contained in N^a or N^c .

Similarly, the absolute value of the $(m_N - |k_1(2q+1) + l_1| + 1)$ th element of $E\{S_{12}\}$ is

$$\begin{aligned} & \left| F \left\{ \sum_{(i_1, j_1) \in \mathcal{C}} y(i_1 - k_1 - j_1 - l_1) e^{j\omega(i_1 - j_1)} \right. \right. \\ & \quad \left. \left. \left[\sum_{(i_2, j_2) \in \mathcal{C}} y(i_2 - k_2 - j_2 - l_2) e^{j\omega(i_2 - j_2)} + D \right] \right\} \right| \\ & \leq \sum_{\substack{(i_1 - k_1 - j_1 - l_1) \in \mathcal{C} \\ (i_2 - k_2 - j_2 - l_2) \in \mathcal{C}}} |F\{y(i_1 - k_1 - j_1 - l_1) e^{j\omega(i_1 - j_1)} \\ & \quad + y(i_2 - k_2 - j_2 - l_2) e^{j\omega(i_2 - j_2)} + D)\}| \\ & \leq 2(m_N + 1)n_c M \end{aligned} \quad (30)$$

Since $e^{j\omega(i_1 - j_1)}$ and $e^{j\omega(i_2 - j_2)}$ have zero means and are uncorrelated with $y(i - k - j - l)$ for $(k, l) \notin \mathcal{N} \cup \{(0, 0)\}$, $e^{j\omega(i_1 - j_1)}$ is uncorrelated with $y(i_2 - k_2 - j_2 - l_2)$ and $e^{j\omega(i_2 - j_2)}$ is uncorrelated with $y(i_1 - k_1 - j_1 - l_1)$ for $(k_1, l_1) \in \mathcal{N}$ and $(k_2, l_2) \in \mathcal{N}$ respectively then

$$|F\{y(i_1 - k_1 - j_1 - l_1) e^{j\omega(i_1 - j_1)} y(i_2 - k_2 - j_2 - l_2) e^{j\omega(i_2 - j_2)}\}| \leq M \quad (31)$$

if $(i_1 - i_2 + k_2 - j_1 - j_2 + l_1) \in \mathcal{N} \cup \{(0, 0)\}$ or $(i_1 - i_2 + k_1 - j_1 - j_2 + l_2) \in \mathcal{N} \cup \{(0, 0)\}$ otherwise the left hand side of (31) equals zero. Thus the absolute value of the $(m_N - |k_1(2q+1) + l_1| + 1 - |k_2(2q+1) + l_2|)$ th element of $F\{S_1\}$ can be expressed as

$$\begin{aligned} & \left| F \left\{ \sum_{(i_1, j_1) \in \mathcal{C}} y(i_1 - k_1 - j_1 - l_1) e^{j\omega(i_1 - j_1)} \right. \right. \\ & \quad \left. \left. \sum_{(i_2, j_2) \in \mathcal{C}} y(i_2 - k_2 - j_2 - l_2) e^{j\omega(i_2 - j_2)} \right\} \right| \\ & \leq \sum_{\substack{(i_1 - k_1 - j_1 - l_1) \in \mathcal{C} \\ (i_2 - k_2 - j_2 - l_2) \in \mathcal{C}}} |F\{y(i_1 - k_1 - j_1 - l_1) e^{j\omega(i_1 - j_1)} \\ & \quad + y(i_2 - k_2 - j_2 - l_2) e^{j\omega(i_2 - j_2)}\}| \\ & \leq 2(m_N + 1)n_c M \end{aligned} \quad (32)$$

and the absolute value of the $(m_N - |k_1(2q+1) + l_1| + 1)$ th element

$$\begin{aligned} & \left| E \left\{ \sum_{(i_1, j_1) \in \mathcal{C}} y(i_1 - k_1 - j_1 - l_1) e^{j\omega(i_1 - j_1)} \right. \right. \\ & \quad \left. \left. \left[\sum_{(i_2, j_2) \in \mathcal{C}} y(i_2 - k_2 - j_2 - l_2) e^{j\omega(i_2 - j_2)} + D \right] \right\} \right| \\ & \leq \sum_{\substack{(i_1 - k_1 - j_1 - l_1) \in \mathcal{C} \\ (i_2 - k_2 - j_2 - l_2) \in \mathcal{C}}} |E\{y(i_1 - k_1 - j_1 - l_1) e^{j\omega(i_1 - j_1)} \\ & \quad + y(i_2 - k_2 - j_2 - l_2) e^{j\omega(i_2 - j_2)} + D)\}| \\ & \leq 2(m_N + 1)n_c M \end{aligned} \quad (33)$$

By a similar procedure we can prove that for $(k_1, l_1), (k_2, l_2) \in \mathcal{N}$ the following statements are true

- 1) the absolute value of $(|k_1(2q+1) + l_1|)$ th element of $F\{S_2\} \leq 2(m_N + 1)n_c M$
- 2) the absolute value of $(|k_2(2q+1) + l_2|)$ th element of $F\{S_1\} \leq 2(m_N + 1)n_c M$
- 3) the absolute value of $(|k_1(2q+1) + l_1| - |k_2(2q+1) + l_2|)$ th element of $E\{S_1\} \leq 2(m_N + 1)n_c M + n_c M$
- 4) $|L\{S_1\}| \leq 2(m_N + 1)n_c M + n_c M$
- 5) $|F\{S_{11}\}| \leq 2(m_N + 1)n_c M + n_c M$ and
- 6) $|F\{S_{41}\}| \leq 2(m_N + 1)n_c M + n_c M$

To summarize the absolute value of any element of $F\{S\}$ does not exceed $O(n_c)$ as $n_c \rightarrow \infty$.

Because the maximum absolute value of all elements of a matrix is a norm of the matrix under the conditions of the above lemma the norm of $F\{R_i(\zeta)[g_i - g_j(\zeta)][g_i - g_j(\zeta)]^T R_i^T(\zeta)\}$ does not exceed $O(n_c)$ as $n_c \rightarrow \infty$.

Theorem Given a stationary image $\{y(i, j) | (i, j) \in B\}$ the modified LS estimator defined by (14) is a consistent estimator if the conditions in the lemma are valid and (34) (found at the bottom of the page) is a positive definite matrix.

Proof For large value of n_c

$$R_i(\zeta) = n_c Q + \Theta(n_c) \quad (35)$$

where Q is defined by (34) and $\Theta(n_c) = R_i(\zeta) - n_c Q$ is a matrix satisfying the condition that

$$F\{\Theta^2(n_c)\} = O(n_c) \quad (36)$$

$$Q = \begin{bmatrix} JF\{x_i^T(i, j)x_i^T(i, j)\}J & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & F\{x_i^T(i, j)x_i^T(i, j)\} \end{bmatrix} \quad (34)$$

Therefore

$$\begin{aligned} & E\{R_p(G)[g_p - g_p(G)][g_p - g_p(G)]^T R_p^T(G)\} \\ &= E\{[n_G Q + \Theta(n_G)][g_p - g_p(G)][g_p - g_p(G)]^T [n_G Q + \Theta(n_G)]^T\} \\ &\simeq n_G^2 Q E\{[g_p - g_p(G)][g_p - g_p(G)]^T\} Q + O(n_G^{1/2}). \end{aligned} \quad (37)$$

The application of the lemma, which says $\|E\{R_p(G)[g_p - g_p(G)][g_p - g_p(G)]^T R_p^T(G)\}\|$ does not exceed $O(n_G)$ as $n_G \rightarrow \infty$, leads to

$$Q E\{[g_p - g_p(G)][g_p - g_p(G)]^T\} Q \rightarrow 0 \quad \text{as } n_G \rightarrow \infty \quad (38)$$

and thus the consistency of the estimate $g_p(G)$ follows because of the positive definiteness of Q .

IV. CONCLUDING REMARKS

In this note, we have proven analytically the consistency of the modified LS estimator recently introduced in [17], which can serve as a very important basis for choosing the best estimator. The conditions for the consistency are 2)–5). There have been some discussions on conditions 1)–3) in [17]. Condition 4) can be valid in practical applications where $\{y(i, j)\}$ is bounded, and therefore $\{e'(i, j)\}$ and $\{e''(i, j)\}$ are bounded. Condition 5) means $e'(i, j)$ and $e''(i, j)$ are uncorrelated with $y(k, l)$ far away from the point (i, j) , which is also very natural in real applications.

Extensive computer simulations with changing size data sample windows have been done. The results of the computer simulations show the following:

- 1) The mean square errors for the modified LS estimator decrease, and the modified LS parameter estimates converge to their corresponding true parameter values with the greatest rate when the sizes of the data sample windows increase.
- 2) The rate of decrease in the mean square error for the modified LS estimator is about seven times greater than that for the approximate ML.
- 3) The conventional LS estimator has the least rate of decrease in its mean square errors because, in this case, the conventional LS is not theoretically a consistent estimator.

The computer simulation methods and results are omitted from this note due to the space limitation, but they are available upon request.

REFERENCES

- [1] J. W. Woods and C. Radewan, "Kalman filtering in two dimensions," *IEEE Trans Inform Theory*, vol. IT-23, pp. 473–482, 1977.
- [2] H. Kaufman, J. W. Woods, S. Dravida, and A. M. Tekalp, "Estimation and identification of two-dimensional images," *IEEE Trans Automat Contr*, vol. AC-28, pp. 745–756, 1983.
- [3] J. W. Woods, S. Dravida, and R. Mediavilla, "Image estimation using doubly stochastic Gaussian random field models," *IEEE Trans Pattern Anal Machine Intel*, vol. PAMI-9, pp. 245–253, 1987.
- [4] F. C. Jeng and J. W. Woods, "Compound Gauss–Markov random fields for image estimation," *IEEE Trans Signal Processing*, vol. 39, pp. 683–697, 1991.
- [5] H. C. Andrews and B. R. Hunt, *Digital Image Restoration*. Englewood Cliffs, NJ: Prentice-Hall, 1977.
- [6] A. K. Jain and J. R. Jain, "Partial differential equations and finite difference methods in image processing, Part II: Image restoration," *IEEE Trans Automat Contr*, vol. AC-23, pp. 817–834, 1978.
- [7] J. W. Woods, "Two-dimensional Markov spectral estimation," *IEEE Trans Inform Theory*, vol. IT-22, pp. 552–559, 1976.
- [8] B. F. McGuffin and B. Liu, "An efficient algorithm for two-dimensional autoregressive spectrum estimation," *IEEE Trans Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 106–117, 1989.
- [9] R. R. Hanser, Jr., and R. Chellappa, "Noncausal 2-D spectrum estimation for direction finding," *IEEE Trans Inform Theory*, vol. 36, pp. 108–125, 1990.
- [10] R. Chellappa and R. L. Kashyap, "Texture synthesis using 2-D noncausal autoregressive models," *IEEE Trans Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 194–203, 1985.
- [11] P. Y. Zhao, A. R. Figueiras-Vidal, and J. R. Casar, "Color texture analysis and synthesis," Rep., Universidad Politecnica de Madrid, Nov 1991.
- [12] P. Y. Zhao and Z. Y. He, "Multidimensional system identification and applications," *Advances Model Simul.*, vol. 18, no. 1, pp. 15–26, 1989.
- [13] P. Y. Zhao and D. R. Yu, *Multidimensional System Identification*. Shanghai, China: Shanghai Scientific & Technical Publishers, 1994 (in Chinese).
- [14] J. W. Woods, "Two-dimensional discrete Markovian fields," *IEEE Trans Inform Theory*, vol. IT-18, pp. 232–240, 1972.
- [15] —, "Markov image modeling," *IEEE Trans Automat Contr*, vol. AC-23, pp. 846–850, 1978.
- [16] R. L. Kashyap and R. Chellappa, "Estimation and choice of neighbors in spatial-interaction models of images," *IEEE Trans Inform Theory*, vol. IT-29, pp. 60–72, 1983.
- [17] P. Y. Zhao and D. R. Yu, "An unbiased and computationally efficient LS estimation method for identifying parameters of 2-D noncausal SAR models," *IEEE Trans Signal Processing*, vol. 41, pp. 849–857, 1993.
- [18] R. L. Kashyap, "Characterization and estimation of two-dimensional ARMA models," *IEEE Trans Inform Theory*, vol. IT-30, pp. 736–745, 1984.
- [19] P. Y. Zhao and D. R. Yu, "Spatially recursive algorithm to estimate 2-D ARMA model parameters," *Sci Lett*, vol. 36, no. 8, pp. 578–581, 1991 (in Chinese).
- [20] A. K. Jain, "Advances in mathematical models for image processing," *Proc IEEE*, vol. 69, pp. 502–528, 1981.
- [21] P. Y. Zhao and Z. Y. He, "A fast spatial recursive algorithm for 2-D adaptive LS filtering and prediction," *Advances Model Simul.*, vol. 13, no. 1, pp. 57–64, 1988.
- [22] P. Y. Zhao and D. R. Yu, "Spatially recursive algorithms for adaptive estimation of two-dimensional non-causal and non-stationary simultaneous autoregressive model parameters," *Int J Syst Sci*, vol. 23, pp. 1033–1049, 1992.
- [23] Y. Uetake, "Realization of noncausal 2-D systems based on a descriptor model," *IEEE Trans Automat Contr*, vol. 37, pp. 1837–1840, 1992.
- [24] Y. Uetake, "Optimal smoothing for noncausal 2-D systems based on a descriptor model," *IEEE Trans Automat Contr*, vol. 37, pp. 1840–1845, 1992.

A Robust Hybrid Stabilization Strategy for Equilibria

John Guckenheimer

Abstract—For an equilibrium of a general dynamical system, the domain of stability of a linear feedback controller is enlarged by the use of a general “hybrid” or “switching” strategy. The strategy is illustrated for numerical simulations of an inverted double pendulum on a cart.

I. INTRODUCTION

Linear control strategies provide a means for the stabilization of equilibria under general hypotheses. When applied to nonlinear systems, the effectiveness of these strategies depends upon the size of the domain of stability that is produced for the stabilized equilibrium. If this domain is small compared to the accuracy of the measurements or the size of disturbances within the system, then the linear controller is likely to fail within a short period. Failure of the system can be catastrophic, with the system wandering far from the desired equilibrium. We present here a general procedure to recapture stability of a linear controller when a trajectory leaves its region of stability. By using a hybrid strategy based upon discrete switching events within the state space of the plant, the system returns to the region of stability for the linear controller from a much larger domain. The control procedure is robust and remains effective under large classes of perturbations of the underlying system. We illustrate the effectiveness of our technique by applying it to the control of an inverted double pendulum.

The stabilization of unstable equilibria is a fundamental problem for the control of engineering systems. A sufficient condition for stability of an equilibrium point of a smooth dynamical system is that the eigenvalues of its linearization lie in the left-half plane. This is easily proved by several means, for example, by defining a quadratic Lyapunov function in a neighborhood of the equilibrium. Control theory addresses the questions of when stabilization is possible with the modifications (controller) that can be built into the underlying system (plant). Over the past 50 years, an extensive theory of “linear control” has developed comprehensive procedures for determining when the stabilization problem is solvable and for the design of controllers that implement stabilization. This theory is widely employed in engineering for the design of controllers in communication systems, chemical process control, avionics, etc. However, linear feedback control is not a complete panacea for all control problems, even ones of stabilizing equilibria when complete state-space information is available at all times. One difficulty that is encountered in some applications is that the domain of attraction of a controlled equilibrium may be small. This leads to unacceptable constraints on system performance. Small random disturbances in the environment or the inability of the actual physical system to implement its model idealization lead to failures of the controller. The results of the failure can be catastrophic in terms of the design objectives. For example, in the double pendulum example we describe below, the failure of a linear controller leads to large motions of the

pendulum, and in the presence of damping, the pendulum eventually falls to rest at its naturally stable hanging position.

The goal of this work is to provide a simple, effective means of recovery from the failure of a linear controller. We want to design a “safety net” around the (small) domain of attraction of a linear controller, so that if a disturbance moves a system outside this domain of attraction, it will be guided back into the domain by the application of a different control strategy. The strategy that we describe is very general. It can be applied to any system that meets conditions of controllability and accessibility. Moreover, the computations that are required for the design of a controller are based on the linearization of the system at its equilibrium, just as with linear controllers. Verification of the effectiveness of a particular design requires more extensive simulation, but the design guidelines are based upon accessible information.

The framework in which our control strategy is implemented has precursors in the literature [2], [3], [11], [12], [15]. The terms “switching” system, “variable structure” system, and “hybrid” system have all been used to describe piecewise smooth vector fields in the context of control, but there does not yet seem to be an effective, systematic theory of such systems. We shall use the terms switching system and hybrid system interchangeably. One of the essential aspects of our work is the presence of “hysteresis” in a piecewise smooth system: there is a discrete component of the state of the system used by the controller in addition to its location in the underlying state space of the physical system. We recall the description of hybrid systems that we have used previously [1] and adopt here.

The problem domain is a disjoint union of open, connected subsets of R^n , called charts. Each chart has associated with it a vector field. Inside each chart is a patch—an open subset with closure contained inside the patch. The patch boundaries are assumed piecewise smooth. The evolution of the system is implemented as a sequence of trajectory segments where the endpoint of one segment is connected to the initial point of the next by a transformation, although the transformations are trivial in the examples studied in this note. However, states of a system have both a continuous and discrete part, and switches that change the discrete part of the system state do occur. Time is divided into contiguous periods separated by instances where a state transformation is applied at times referred to as events.

We end this section with a few “philosophical” comments underlying our approach to nonlinear control. Structural stability is a useful concept for dynamical systems, stating that perturbations of a system remain equivalent to the reference system by continuous changes of coordinates. In implementing hybrid control for stabilizing equilibria, we have sought to maintain this type of robustness to perturbations of the system itself. Nothing in the controller should be subject to the choice of exact values of any parameters. In particular, we have avoided the use of sliding modes or switches that must be implemented exactly to be effective. To a large extent, this strategy involves trade-offs to attempts to achieve optimization of some cost function in the control because that is likely to push one to select parameter values in a system that are borderline for a property. If the property is one that involves the stability of the system, then small perturbations cannot be relied upon to maintain the efficacy of the controller. Here the emphasis is squarely upon reliability rather than efficiency. It is quite possible that modifications of the strategy described here will lead to improved performance by returning a system to the stability region of a linear controller more quickly.

Manuscript received January, 1993, revised August 6, 1993 and May 23, 1994. This work was supported by ARPA under U.S. Army Contract DAA21-92-C-0013 to ORA Corporation, administered by ARDEC at the Picatinny Arsenal; and in part by the Air Force Office of Scientific Research under Grant F49620-92-J-0287, the National Science Foundation, and the Department of Energy under Grant DE-FG02-93ER25164.

The author is with the Center for Applied Mathematics, Cornell University, Ithaca, NY 14853 USA.

IEEE Log Number 9406995.

II. BARRIERS TO UNBOUNDED MOTION

This section discusses a strategy for maintaining the motion of a trajectory in a bounded region of an unstable equilibrium point with a piecewise constant control. Consider the following linear system as a model example:

$$\begin{aligned}\dot{x}_1 &= \lambda_1(x_1 - c) \\ \dot{x}_2 &= \lambda_2(x_2 - c).\end{aligned}\quad (1)$$

In (1), c represents a "control" that moves the equilibrium point of the systems along a line. While we have chosen a particular form for this system, most planar vector fields with real eigenvalues can be transformed to this representation by a linear change of coordinates. Such a transformation to a "normal form" exists if the system is controllable. Necessary and sufficient conditions for this are that 1) the eigenvalues are distinct and 2) the control moves the equilibrium along a line that is not an eigendirection. We assume that $\lambda_1 > \lambda_2 > 0$, so that equilibrium point is a source.

The goal is to define a feedback control $c(x)$ so that the motion of the system remains within a moderate-sized bounded neighborhood of the origin. We approach this by defining a hybrid system with two patches that are half-planes H_+ and H_- defined by $y > x - c$, $y < x + c$, respectively. The boundaries of H_+ and H_- are the lines L_{\pm} defined by $y = x \mp c$. These lines are parallel to the control line. In each of the two patches, the control $c(x)$ takes a constant value c_{\pm} . The values of c are chosen with the object of making trajectories in the overlap strip $H_+ \cap H_-$ stay in a bounded region of the origin. This defines a hybrid system with parameters c_+ and c_- . The goal is now clear: to choose values of these parameters to create a trapping region surrounding the origin.

We can compute readily that the trajectories of the system (1) are defined by

$$\begin{aligned}x_1(t) &= c + \exp(t\lambda_1)(x_1(0) - c) \\ x_2(t) &= c + \exp(t\lambda_2)(x_2(0) - c).\end{aligned}$$

Given c , we would like to find c_{\pm} that creates a trapping region around the origin in the strip $-c \leq x_2 - x_1 < c$. Along a segment of the right boundary L_+ of the strip, we would like the vector field associated with H_+ to point to the left, toward the interior of the strip. Similarly, we would like the vector field associated with H_- to point to the right on a segment of the left boundary L_- of H_- . These segments are to be chosen so that the flow carries one into the other. See Fig. 1. To choose c_{\pm} with the desired properties, we argue as follows. For simplicity, we shall assume that $c_+ = -c_-$ so that the system has a symmetry. The symmetry streamlines the analysis, but is not essential to the argument that we give.

Regard the value of c as fixed for the moment. We shall determine the conditions we would like it to satisfy. Along L_+ , define the point p_+ to be the point where the vector field has slope 1. The point p_+ is obtained by solving the equations

$$\begin{aligned}x_2 &= x_1 - c \\ \lambda_1(x_1 - c) &= \lambda_2(x_2 - c)\end{aligned}$$

whose solution is

$$(x_1, x_2) = \left(c - \frac{\lambda_2}{\lambda_1 - \lambda_2}c, c - \frac{\lambda_1}{\lambda_1 - \lambda_2}c \right).$$

Above p_+ on L_+ , the vector field points to the right of L_+ . Below p_+ on L_+ , the vector field points to the left of L_+ . The trajectory starting at p_+ should lie above the trapping region in the strip. For example, we might want the intersection of the $x_2 - x_1$ axis with the strip to lie in the trapping region. For this to occur, it suffices that the trajectory with initial condition p_+ intersect the strip at a point

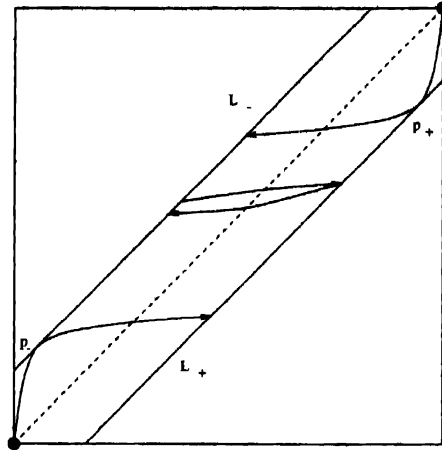


Fig. 1. The barriers L_{\pm} and region in which trajectories switch back and forth between the boundaries. The curves segments show trajectories, with arrows located where they encounter a patch boundary

TABLE 1

λ	c/e
1.5	9.444
2.0	5.828
2.5	4.614
3.0	4.0
3.5	3.629
5.0	3.063
10	2.518
100	2.065

with a nonnegative value of x_1 . A lower bound for c_+ satisfying this criterion is given by the value of c for which the trajectory with initial condition p_+ passes through the point $(0, c)$. This yields an implicit equation for c/e in terms of the ratio $\lambda = \lambda_1/\lambda_2$

$$\frac{\left(\frac{c}{e}\right)^{\lambda_1}}{\left(\frac{c-c}{e}\right)^{\lambda_2}} = \frac{(\lambda_1 - \lambda_2)^{(\lambda_1 - \lambda_2)}(\lambda_2)^{\lambda_2}}{\lambda_1^{\lambda_1}}.$$

Representative values of the solution of this equation are given in Table I to three-digit accuracy. This discussion leads to the following theorem.

Theorem 2.1 Let c_+/c be larger than the solution of the equation

$$\frac{\left(\frac{c}{e}\right)^{\lambda_1}}{\left(\frac{c-c}{e}\right)^{\lambda_2}} = \frac{(\lambda_1 - \lambda_2)^{(\lambda_1 - \lambda_2)}(\lambda_2)^{\lambda_2}}{\lambda_1^{\lambda_1}}$$

and set $c_- = -c_+$. Denote by W_{\pm} the trajectories of the vector fields X_{\pm} defined by

$$\begin{aligned}\dot{x}_1 &= \lambda_1(x_1 - c_{\pm}) \\ \dot{x}_2 &= \lambda_2(x_2 - c_{\pm})\end{aligned}$$

passing through the points $(0, \pm c)$. Assume $\lambda_2 < \lambda_1$. On the lines l_{\pm} defined by $x_2 = x_1 \mp c$, consider the segment s_{\pm} with endpoints at $(0, \mp c)$ and the intersections of W_{\pm} with l_{\pm} . Then trajectories of X_+ with initial conditions on l_+ intersect l_- , and trajectories of X_- with initial conditions on l_- intersect l_+ .

To prove this theorem, we still need to verify that the trajectory of X_- with initial condition at $(0, c)$ intersects the segment s_+ . From X_- , we obtain the equation

$$\frac{dx_2}{dx_1} = \frac{\lambda_2(x_2 + c)}{\lambda_1(x_1 + c)} \quad (2)$$

Equation (2) gives the trajectory of X_- with initial condition $(0, c)$ in a form parameterized by x_1 :

$$\left(x_1, -c + (c + \epsilon)\left(1 + \frac{x_1}{c}\right)^\lambda\right)$$

with $\lambda = \lambda_2/\lambda_1$. Similarly, the trajectory of X_+ with initial condition $(0, c)$ is given by

$$\left(x_1, c + (c - \epsilon)\left(1 - \frac{x_1}{c}\right)^\lambda\right).$$

These two trajectories intersect at one point in the right-half plane if the trajectory of X_- passes below the point (c, c) . Setting $u = x_1/c$ and $b = c/\epsilon$, we want to determine when

$$-1 + (1 + b)(1 + u)^\lambda = 1 - (1 - b)(1 - u)^\lambda.$$

This equation is readily solved for b in terms of u

$$b = \frac{2 - (1 - u)^\lambda - (1 + u)^\lambda}{(1 + u)^\lambda - (1 - u)^\lambda}.$$

At the solution to this equation, we want to have $-1 + (1 + b)(1 + u)^\lambda - u + b > 0$, so that the intersection point of the two trajectories lies above the line l_+ . This yields the requirement that

$$b > \frac{(1 + u) - (1 + u)^\lambda}{(1 + u) - (1 + u)^\lambda}.$$

Substituting the value for b at the intersection point into this inequality and simplifying leads to the requirement that

$$2 - u(1 + u)^\lambda + u(1 - u)^\lambda - 2(1 - u^2) > 0.$$

This inequality holds throughout the unit square in the (u, l) plane, which is the domain of interest. We conclude that the intersection of the trajectories of X_+ and X_- with common initial point $(0, c)$ lies above the line l_+ since we assumed $\lambda < 1$.

Recall that our hybrid system X applies a mode switch from X_- to X_+ when a trajectory hits l_+ from the left and, similarly, a mode switch from X_+ to X_- when a trajectory hits l_- from the right. The theorem implies the following corollary.

Corollary 2.1: With the notations introduced above, define R to be the region bounded by W_\pm , the trajectories of X_\pm with initial conditions at $(0, \mp c)$ and s_\pm . Then trajectories with initial conditions in R remain in R for all forward time.

Less formally, we say that R is a trapping region for the hybrid system X . We can say more still about the dynamics of X in R . There are passage maps θ_+ that map trajectories with initial conditions on s_\pm to their intersections with s_+ . The maps θ_\pm are monotone, so the composition $\theta_- \circ \theta_+$ is a monotone map of the interval s_+ into itself. It follows that this return map has a stable fixed point, representing a stable periodic orbit for the hybrid system. Additional computations lead to the conclusion that the return map is a contraction and has only a single fixed point. To carry forward these computations, we use the coordinates which scale c to 1, writing $u = x_1/c$ and $b = c/\epsilon$ as in the proof of the theorem. The trajectory starting on the line l_- with initial value $u = u_0$ is given by

$$\left(u_0, -1 + (1 + b + u_0)\left(\frac{1 + u}{1 + u_0}\right)^\lambda\right).$$

The intersection of the trajectory occurs at a value of $u = u_1$ satisfying

$$-1 + (1 + b + u_0)\left(\frac{1 + u}{1 + u_0}\right)^\lambda - u_1 + b = 0.$$

We want to estimate du_1/du_0 from this equation. From the equation

$$\frac{(1 + u_0)^\lambda}{1 + u_0 + b} = \frac{(1 + u_1)^\lambda}{1 + u_1 - b}$$

we deduce that $du_1/du_0 < 1$. The right-hand side of the last equation defines a function of u_1 which intersects the function of u_0 on the left-hand side crossing from above to below while decreasing. Therefore implicitly differentiating the last equation gives $du_1/du_0 < 1$. From this, we conclude that the return map of our hybrid system has derivative smaller than 1 and is a contraction.

Theorem 2.2: With the same hypotheses of the previous theorem and corollary, there is a stable limit cycle for the hybrid system that is globally attracting for all initial conditions in the trapping region R .

Remarks:

1) Due to the symmetry of the hybrid system, the stable limit cycle is symmetric with respect to the origin.

2) In systems with equilibria that are saddle points with two-dimensional unstable manifolds, the procedure described above can be applied with switching surfaces that are hyperplanes tangent to the directions spanned by the stable manifold of the equilibrium and the tangent to the control curve. We have not investigated systems with unstable manifolds of dimension larger than two, but the strategy might work there as well. In that case, one can try to use switching surfaces that are tangent to the directions spanned by the tangent to the control curve and all but the largest two eigendirections.

III. MULTIPLE BARRIERS

The results of the previous section can be extended and improved in a number of ways. We describe two.

The barriers described in the previous section can be combined with linear controllers. If one knows a region around the equilibrium that lies in the domain for a linear controller, then one can define a hybrid system with three patches: the system described in the previous section, and a domain cut from these two patches in which the linear controller will be applied. If the stable limit cycle of the switching system intersects the domain in which the linear controller is applied, then the barriers and switching system serve to guide the system to the domain of the linear controller from initial conditions between the two barriers. To prevent the system from exiting the domain of the linear controller, distinct boundaries can be defined to switch the linear controller on and off.

The second extension of the controller described in the previous section is to place multiple barriers in the system parallel to one another. Consider, for example, a planar system with six barriers that are parallel lines l_i , $i = 1, \dots, 6$. The lines l_i divide the plane into seven closed "strips" S_i , $i = 1, \dots, 7$. S_1 and S_7 are half-planes. From the S_i , we form six overlapping patches $D_i = S_i \cup S_{i+1}$. In each of these patches, define a constant control that increases in magnitude as one moves away from the control line. The transition conditions are defined so that if one crosses a patch boundary moving away from the control line, then the control setting of larger magnitude is applied. If one crosses a patch boundary moving toward the control line, then the control value changes to one of smaller magnitude. The effect of these barriers is to guide a trajectory back toward the origin from farther away from the origin, while at the same time decreasing the amplitude of the control when feasible. Combining these multiple barriers with a linear controller in a neighborhood of the origin allows one to recover from disturbances of large size that move the system outside the region of stability for the linear controller. See Fig. 2 for an illustration of the patches

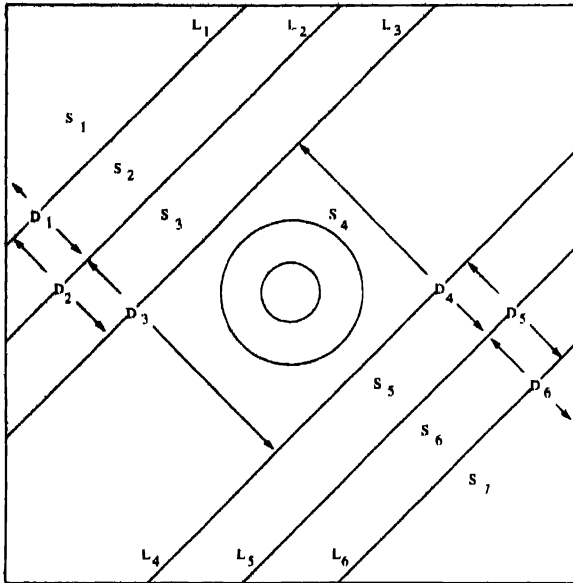


Fig. 2. The geometry of the state space for the system with three sets of barriers and a region in which linear feedback control is used.

associated with a system that has three pairs of barriers and a domain where the linear controller will be applied.

IV. AN EXAMPLE: THE DOUBLE PENDULUM ON A CART

We describe an example—a frictionless double pendulum on a zero mass cart whose acceleration along a track can be controlled. The object is to keep the pendulum in the fully upright position. The control of a pendulum on a cart has been a frequently studied problem [4]–[10], [13], [14]. This example provides a good illustration of the effectiveness of our stabilization strategy on a nonlinear system.

The double pendulum consists of two point masses m_1 and m_2 with body 1 attached to a fulcrum and body 1 attached to body 2 by massless rigid rods of lengths l_1 and l_2 . We want to include within the system of equations the additional effect of applying a horizontal acceleration. See Fig. 3. Choose units for which the acceleration of gravity is 1, let the magnitude of the acceleration be α , and set $\mu = 1 + m_1/m_2$. Then the equations of motion are given by the following vector field X as shown in the equation at the bottom of the page. Here, q_1, q_2 are angular coordinates and p_1, p_2 are the conjugate momenta. The angles q_1, q_2 are measured with respect to vertical rays pointing down, so the stable equilibrium with the pendulum hanging down is given by $q_1 = q_2 = p_1 = p_2 = 0$. The vertically upright position that we want to stabilize is given by $q_1 = q_2 = \pi$ and $p_1 = p_2 = 0$.

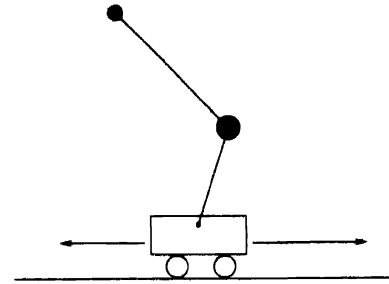


Fig. 3. A double pendulum on a cart.

The vertically upright position is an equilibrium of the pendulum equations (without horizontal acceleration) that has a two-dimensional stable manifold and a two-dimensional unstable manifold. Therefore, we are in a situation for which the theory described earlier can be applied. To do so, we need to construct a linear controller, a region in which the linear controller will be applied, barriers parallel to the hyperplanes spanned by the control line and the stable manifold at the vertical equilibrium, and control values for each of the patches to be used by the controller outside the patch of the linear controller.

The linear controller is defined by making the acceleration of the pendulum fulcrum a linear function of the location of the pendulum in phase space. For convenience, we shall use coordinates $(-\sin(q_1), -\sin(q_2), p_1, p_2)$ near the upright equilibrium. We seek a vector $\gamma = (g_1, g_2, q_1, q_2)$ so that setting $\alpha = g_1 q_1 + g_2 q_2 + g_1 p_1 + g_2 p_2$ makes the upright equilibrium stable. Controllability of the system linearized at the upright equilibrium implies that we can find γ to place the eigenvalues of the linearly controlled system anywhere in the complex plane. We describe one approach to solving this problem. Treating the eigenvalues of the linearization as functions of the control coefficients g_i gives a system of equations that can be solved for the g_i . Let $\lambda_i, i = 1 \dots 4$ be the desired eigenvalues for the controlled system. Denoting the Jacobian matrix of the vector field by A and $v = \partial A / \partial \alpha$, we seek vectors $u_i \neq 0$ and γ so that

$$(A + \gamma \cdot v^T)u_i = \lambda_i u_i.$$

If $A + \lambda_i I$ is invertible, rewrite this equation as

$$-\gamma(u_i)(A - \lambda_i I)^{-1}v = u_i.$$

It follows that u_i is a multiple of $(A - \lambda_i I)^{-1}v$ and

$$\gamma((A - \lambda_i I)^{-1}v) = -1.$$

As i varies in $\{1, 2, 3, 4\}$, this yields a system of linear equations for γ . If the system is nonsingular, then it uniquely determines γ in terms of the eigenvalues λ_i .

We have investigated a numerical example with parameters $l_1 = 1/2, l_2 = 3/4, m_1 = 2, m_2 = 1$, and

$$\begin{aligned} \dot{q}_1 &= p_1 \\ \dot{q}_2 &= p_2 \\ \dot{p}_1 &= \frac{\sin(q_2) \cos(q_1 - q_2) - \mu \sin(q_1) - (l_2 p_2^2 + l_1 p_1^2 \cos(q_1 - q_2)) \sin(q_1 - q_2)}{l_1(\mu - \cos^2(q_1 - q_2))} \\ &\quad + \frac{\alpha(-\mu \cos(q_1) + \cos(q_2) \cos(q_1 - q_2))}{l_1(\mu - \cos^2(q_1 - q_2))} \\ \dot{p}_2 &= \frac{\mu(\sin(q_1) \cos(q_1 - q_2) - \sin(q_2)) + (\mu l_1 p_1^2 + l_2 p_2^2 \cos(q_1 - q_2)) \sin(q_1 - q_2)}{l_2(\mu - \cos^2(q_1 - q_2))} \\ &\quad + \frac{\alpha(-\mu \cos(q_2) + \mu \cos(q_1) \cos(q_1 - q_2))}{l_2(\mu - \cos^2(q_1 - q_2))}. \end{aligned}$$

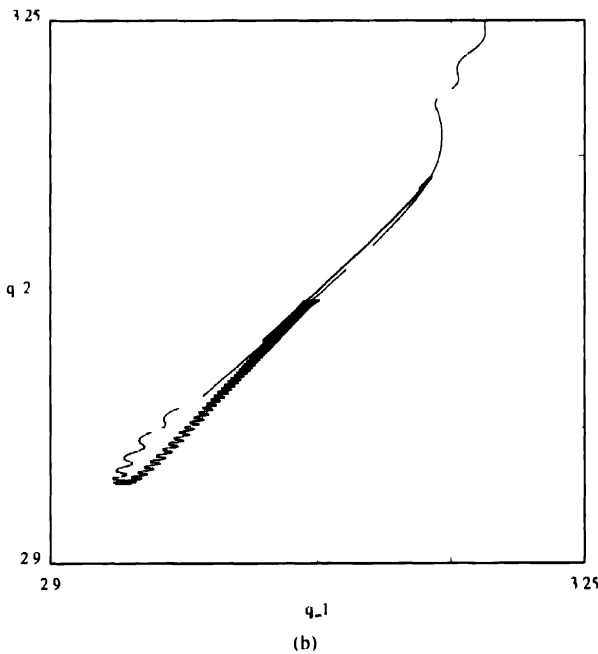
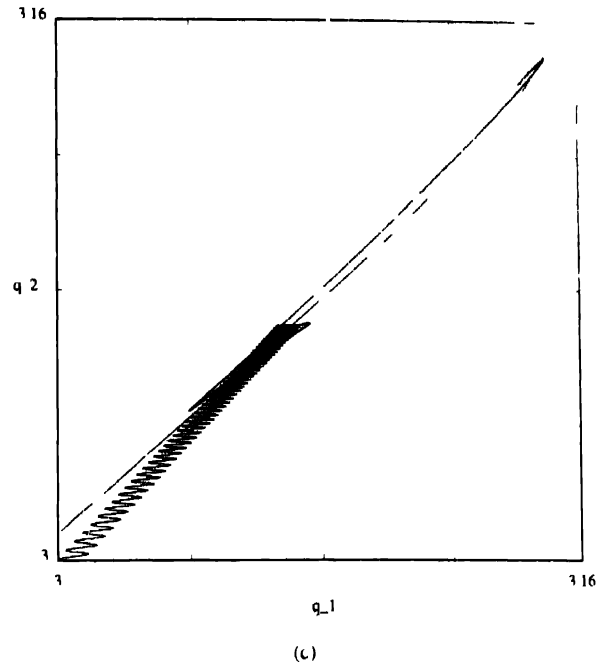
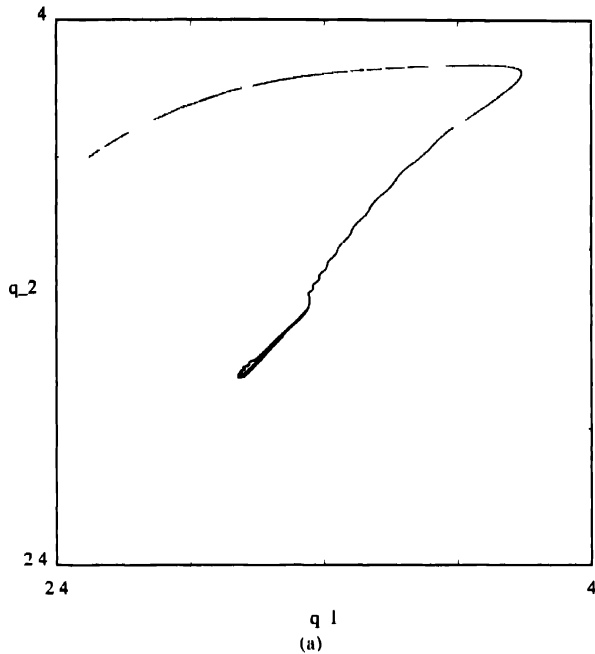


Fig 4 Continued

Fig 4 A typical hybrid trajectory converging to the upright position with magnifications showing increasing detail near the equilibrium position. The initial point of the trajectory is (2.5 3.6) near the left side of (a).

$\{\lambda_1 \lambda_2 \lambda_3 \lambda_4\} = \{-0.1 -0.5 -0.6 -0.7\}$ This gives (exactly) $\gamma = (-3.395 \ 2.416 \ -1.1 \ 1.2595)$. Using these parameter values, we investigated numerically a model of the double pendulum with the hybrid switching control strategy described in the last section. The only modification of the strategy was to define two neighborhoods of the vertical pendulum on which switches to and from the linear controller were applied. This was done because many of the linearly controlled trajectories do not approach the equilibrium with monotonically decreasing distance.

To test the effectiveness of this controller, trajectories were computed on a grid of initial conditions in the plane $p_1 = p_2 = 0$. The domain of stability of the linear controller in this plane contains

a small elongated region around the equilibrium diagonal $q_1 = q_2$ whose length along the diagonal is approximately 0.5 and whose width is approximately 0.07. For the switching system we used level sets of the function $h = (q_1 - \pi)/4 - (q_2 - \pi)/4 + p_1/12 - p_2/6$ as the switching surfaces. The function h was computed so that it is parallel to the hyperplane spanned by the control direction $(0 \ 0 \ 2 \ 0)$ and the stable eigenvalue at the fully upright equilibrium. With switching surfaces given by $h = -0.5$, $h = -0.3$, $h = -0.1$, $h = 0.1$, $h = 0.3$ and $h = 0.5$ with corresponding values for α of $0.3 \ 1 \ -1 \ 3 \ 0$ there is a much larger domain of attraction for the upright pendulum. In the plane $p_1 = p_2 = 0$ it appears that the square with vertices at the points $(q_1 \ q_2) = (2 \ 2.5)$ and $(q_1 \ q_2) = (3.7 \ 3.7)$ is completely contained in the domain of attraction for the upright pendulum. See Fig 4(a) for a projection of a typically trajectory into the plane $p_1 = p_2 = 0$. Fig 4(b) and (c) shows magnifications of this trajectory close to the upright equilibrium. The width of this square is an order of magnitude larger than the width of the domain of stability for the linear controller in the plane $p_1 = p_2 = 0$. If the linear controller is not used then the asymptotic state is the limit cycle described in the first section. Switching to and from the linear controller when the square of the distance to the upright equilibrium is 0.002 and 0.005 respectively appears to robustly stabilize the pendulum at the precise upright state. Note that $0.005 \approx (0.07)^2$ so that the disk for switching the linear controller off could not be chosen smaller and still remain in the basin of attraction of the vertical equilibrium for the linear controller. Addition of stochastic perturbations to the vector field does not appear to significantly diminish the size of the domain of attraction for the upright equilibrium.

As illustrated by this example the hybrid or switching strategy that we have presented for the stabilization of equilibria appears to be robust. All aspects of the strategy seem to be structurally stable and persistent with respect to very general types of perturbations. It augments linear control for stabilizing equilibria by guiding a trajectory back into the domain of attraction for a linear controller from a much larger region.

ACKNOWLEDGMENT

A. Back has helped with the computing and has invested long hours in developing the "hybrid" version of the program DsTool that was used in our numerical investigations. I would like to thank D. Koditschek for bringing to my attention both the need for universal "nonlinear" control strategies and the apparent effectiveness of strategies based upon discrete events. D. Delchamps has been patient and persistent in guiding me toward control theory literature that bears upon the techniques described in this note. S. Johnson has made helpful comments.

REFERENCES

- [1] A. Back, J. Guckenheimer, and M. Myers, "A dynamical simulation facility for hybrid systems," MSI Tech. Rep. 92-6, 1992.
- [2] R. A. DeCarlo, S. H. Zak, and G. P. Matthes, "Variable structure control of nonlinear multivariable systems: A Tutorial," *Proc. IEEE*, vol. 76, Mar. 1988.
- [3] X. Feng and K. Laporo, "Chaotic motion and its probabilistic description in a family of two dimensional nonlinear systems with hysteresis," *J. Nonlinear Sci.*, vol. 2, pp. 417-452, 1992.
- [4] K. Furuta, H. Kajiwara, and K. Kosuge, "Digital control of a double inverted pendulum on a inclined rail," *Int. J. Contr.*, vol. 32, no. 5, pp. 907-924, 1980.
- [5] K. Furuta, T. Okutani, and H. Sone, "Computer control of a double inverted pendulum," *Comput. Elec. Eng.*, vol. 5, pp. 67-84, 1978.
- [6] D. T. Higdon, "Automatic control of inherently unstable systems with bounded control inputs," thesis, Stanford Univ., 1963.
- [7] M. Henders and A. Soudack, "In-the-large" behavior of an inverted pendulum with linear stabilization," *Int. J. Non-linear Mech.*, vol. 27, pp. 129-138, 1992.
- [8] J. F. Schaefer, "On the bounded control of some unstable mechanical systems," thesis, Stanford Univ., 1965.
- [9] J. F. Schaefer and R. H. Cannon, "On the control of unstable mechanical systems, in *Automat. Remote Contr. III, Proc. 3rd Int. Fed. Automat. Contr. (IFAC)*, vol. 1, 6C.1-6C.13, 1967.
- [10] —, "On the control of unstable multiple-output mechanical systems," *ASME*, pp. 2-12, 19XX.
- [11] T. Seidman, *Switching Systems I*, book in manuscript, 1983.
- [12] —, "Switching systems and periodicity," in *Proc. Symp. Nonlinear semigroups, partial differential equations, and attractors*, Washington, DC, Aug. 3-7, 1987, T. L. Gill, W. W. Zachary, Eds., Lecture notes in mathematics 1394.
- [13] T. Shinbrot, C. Grebogi, and J. Wisdom, "Chaos in a double pendulum," *Amer. J. Phys.*, vol. 60, pp. 491-499, 1992.
- [14] W. R. Sturgen and M. V. Loscutoff, "Application of model control and dynamic observers to control a double pendulum," *Proc. JACC*, pp. 857-865, 1972.
- [15] V. I. Utkin, *Sliding Modes and Their Application in Variable Structure Systems*. USSR: MIR, 1978.

Adaptive Control of Systems with Unknown Output Backlash

Gang Tao and Petar V. Kokotović

Abstract—Adaptive control schemes for systems with unknown backlash at the plant output are developed. In the case of known backlash, a backlash inverse controller guarantees exact output tracking. When the backlash characteristics are unknown, adaptive laws are designed to update the controller parameters and to guarantee bounded input-output stability. Simulations show significant improvements of the system performance achieved by such adaptive backlash inverse controllers.

I. INTRODUCTION

Backlash is common in many components of control systems, such as actuators, sensors, and mechanical connections. A typical backlash example is the mechanical motion due to the imperfect contact of two gears. From the early days of classical control theory, the backlash nonlinearity has been recognized as one of the factors severely limiting the performance of feedback systems by causing delays, oscillations, and inaccuracy.

The backlash characteristic is a nondifferentiable nonlinearity which is often poorly known. Therefore, the control of systems with unknown backlash is an open theoretical problem of major relevance to applications. In [1] and [2], we proposed an adaptive control scheme for systems with unknown backlash at input of the plant, that is, in the actuator. In this note, we address the problem with unknown backlash at the plant output, that is, in the sensor, as shown in Fig. 1. Perhaps the most common example is a position servo: the block $G(D)$ represents the power amplifier motor, and the backlash is in the position sensor such as a potentiometer connected to the motor shaft through a gear box.

A contribution of this note is the construction of new adaptive controller structures for the output backlash problem which is essentially different from the actuator backlash problem. Our new controller structures can be initialized to achieve exact output matching when the plant is known. They lead to a linear parameterization of the closed-loop plant when the backlash is unknown. Such a linear parameterization is crucial for the development of an adaptive scheme to deal with the unknown backlash. Our approach is to develop an adaptive backlash inverse to cancel the unknown backlash effect. A feedback-feedforward controller structure is then combined with such an adaptive backlash inverse to achieve the desired tracking performance.

The note is organized as follows. In Section II, we present the model of the backlash at the output of a linear part and formulate the control problem. In Section III, assuming that the backlash is known, we present a backlash inverse and introduce the idea of backlash inverse control. In Section IV, we develop two adaptive backlash inverse controller structures when the backlash is unknown: one for the linear part known, and the other for the linear part unknown.

Manuscript received April 5, 1993; revised January 28, 1994. This work was supported by the National Science Foundation under Grants ECS-9203491 and ECS-9307545, by the Air Force Office of Scientific Research under Grant F49620-92-J-0495, and by a Ford Motor Company grant.

G. Tao is with the Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903 USA.

P. V. Kokotović is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA.

IEEE Log Number 9406996.

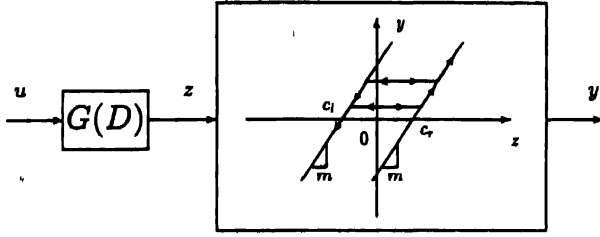


Fig. 1. Plant with output backlash.

We design adaptive laws to update the controller parameters, which ensure closed-loop signal boundedness. In Section V, we present two design examples and use simulation results to show major performance improvements achieved with our adaptive backlash inverse controllers.

II. PROBLEM STATEMENT

Let us consider the following discrete-time plant with a linear part $G(D)$ and a backlash nonlinearity $B(\cdot)$ at its output, as shown in Fig. 1:

$$y(t) = B(z(t)), \quad z(t) = G(D)[u](t) \quad (2.1)$$

where $u(t)$ is the control input, $y(t)$ is the measured output, $G(D) = k_p \frac{Z(D)}{P(D)}$, k_p is a constant, $Z(D)$ and $P(D)$ are monic polynomials, and the symbol D is used to denote, as the case may be, the z -transform variable or the advance operator: $D[x](t) = x(t+1)$.

The backlash characteristic $B(\cdot)$ with input $z(t)$ and output $y(t)$ is described by

$$y(t) = \begin{cases} m(z(t) - c_l) & \text{for } z(t) \leq z_l \\ m(z(t) - c_r) & \text{for } z(t) \geq z_r \\ y(t-1) & \text{for } z_l < z(t) < z_r \end{cases} \quad (2.2)$$

where $m > 0$, c_r, c_l are constant and $z_l = \frac{y(t-1)}{m} + c_l$, $z_r = \frac{y(t-1)}{m} + c_r$ are the z -axis values of the intersections of the two parallel lines of slope m with the horizontal inner segment containing $y(t-1)$. This backlash model is the discrete-time counterpart of the continuous-time backlash characteristic modeled in [1] and [2].

In our control problems, the backlash parameters m, c_r, c_l are unknown, and the internal signal $z(t)$ is not available for measurement. The control objective is to design a feedback control for the plant (2.1), which stabilizes the closed-loop system and makes the plant output $y(t)$ track a given bounded reference signal $y_m(t)$.

We will consider two designs: one for the plant with a known linear part $G(D)$ and the other for an unknown $G(D)$. We will use the model reference control strategy and therefore make the following assumptions: 1) $G(D)$ is minimum phase; 2) the relative degree n^* of $G(D)$ is known; 3) the degree n of $P(D)$ is known; 4) the sign of k_p is known; 5) $\frac{1}{m} \geq m_0$ for some known constant $m_0 \geq 0$, $c_r \geq 0$, $c_l \leq 0$. Assumption 5) will be used to project the estimate of $\frac{1}{m}$ away from zero. Without loss of generality, in view of 4) and 5), we further assume that $k_p = 1$.

Our approach will be to estimate the backlash parameters m, c_r, c_l so that an adaptive backlash inverse can be implemented to cancel the backlash effects. With such an adaptive backlash inverse, a linear controller structure will be designed to generate the control signal $u(t)$ to achieve the control objective.

III. NONADAPTIVE BACKLASH INVERSE CONTROLLERS

In this section, we introduce the idea of backlash inverse control and develop a backlash inverse control scheme when both the linear part $G(D)$ and the backlash $B(\cdot)$ are known.

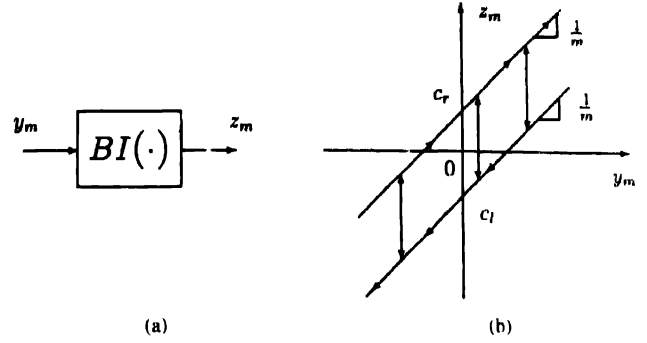


Fig. 2. Backlash inverse: (a) block diagram; (b) characteristic

A Backlash Inverse Control

A graphical inverse of the backlash characteristic in Fig. 1 is shown in Fig. 2, which contains vertical jumps. This backlash inverse will be described with the help of the indicator function $\chi[X]$ of the event X :

$$\chi[X] = \begin{cases} 1 & \text{if } X \text{ is true} \\ 0 & \text{otherwise.} \end{cases}$$

To characterize the upward and downward changes of the reference signal $y_m(t)$, we define

$$\chi_{im}(t) = \chi[y_m(t) > y_m(t-1)] \quad \text{or } y_m(t) = y_m(t-1), \chi_{im}(t-1) = 1] \quad (3.1)$$

$$\chi_{lm}(t) = \chi[y_m(t) < y_m(t-1)] \quad \text{or } y_m(t) = y_m(t-1), \chi_{lm}(t-1) = 1] \quad (3.2)$$

and require that $\chi_{im}(t)$ and $\chi_{lm}(t)$ be initialized as $\chi_{im}(t_0) + \chi_{lm}(t_0) = 1$, $\chi_{im}(t_0)\chi_{lm}(t_0) = 0$. A consequence of this definition is that $\chi_{im}(t) + \chi_{lm}(t) = 1$ and $\chi_{im}(t)\chi_{lm}(t) = 0$ for any $t \geq t_0$.

The output of the backlash inverse in Fig. 2 is

$$z_m(t) = \frac{y_m(t)}{m} + \chi_{im}(t)c_r + \chi_{lm}(t)c_l. \quad (3.3)$$

The mapping $BI(\cdot): y_m(t) \rightarrow z_m(t)$ defines a backlash inverse [1]

$$B(BI(y_m(\tau))) = y_m(\tau) \Rightarrow B(BI(y_m(t))) = y_m(t), \quad \forall t \geq \tau.$$

To introduce the idea of an output matching control law with a backlash inverse, we first consider the simplest case when $G(D)$ in Fig. 1 is an n^* -step delay, that is, $z(t+n^*) = u(t)$.

Proposition 3.1 [3]: Suppose that $G(D) = D^{-n^*}$. Then the control law $u(t) = z_m(t+n^*)$, $t \geq t_0$, guarantees that $y(t+n^*) = y_m(t+n^*)$ for any $t \geq t_0$, provided that $y_m(t_0+n^*) = y(t_0+n^*)$.

In this proposition, the condition that $y_m(t_0+n^*) = y(t_0+n^*)$ is used to initialize the backlash inverse controller for output matching, which can be practically achieved by redefining the initial value of $y_m(t)$. For this simple $G(D)$ and with the backlash characteristic known, the measurement of $z(t)$ is not required.

Next we consider $G(D)$ of a general form with known backlash $B(\cdot)$, but we now assume that the signal $z(t)$ is available for measurement. This unrealistic assumption is made only to help us introduce a nonadaptive controller whose structure will be of interest in our adaptive designs in which this assumption will be dropped.

Using the notation

$$\omega_a(t) = a(D)[u](t), \quad a(D) = (D^{-n^*+1}, \dots, D^{-1})^T$$

$$\omega_b(t) = b(D)[z](t), \quad b(D) = (D^{-n^*+1}, \dots, D^{-1}, 1)^T$$

we choose the parameters $\theta_u^* \in R^{n-1}$ and $\theta_z^* \in R^n$ to satisfy the matching equation

$$\theta_u^{*T} a(D)P(D) + \theta_z^{*T} b(D)Z(D) = P(D) - Z(D)D^{n^*} \quad (3.4)$$

and define our nonadaptive controller as

$$u(t) = \theta_u^{*T} \omega_u(t) + \theta_z^{*T} \omega_z(t) + z_m(t + n^*), \quad t \geq t_0. \quad (3.5)$$

Proposition 3.2 [3]: The control law (3.5) guarantees that $z(t + n^*) = z_m(t + n^*)$ for any $t \geq t_0$. Moreover, $y(t + n^*) = y_m(t + n^*)$ for any $t > t_0$, provided that $y_m(t_0 + n^*) = y(t_0 + n^*)$.

Our designs in subsequent sections will avoid the unrealistic assumption that the signal $z(t)$ is available for measurement by replacing the term $\theta_z^{*T} \omega_z(t)$ in (3.5) with terms implementable without the knowledge of $z(t)$.

B. A Backlash Inverse Controller for $G(D)$ and $B(\cdot)$ Known

To develop a controller structure which uses only the measured plant output $y(t)$, we define

$$\lambda_+(t) = \lambda[y(t) > y(t-1) \text{ or } y(t) = y(t-1), \lambda_+(t-1) = 1] \quad (3.6)$$

$$\lambda_-(t) = \lambda[y(t) < y(t-1) \text{ or } y(t) = y(t-1), \lambda_-(t-1) = 1] \quad (3.7)$$

$$\lambda_0(t) = \lambda[y(t) = y(t-1)]. \quad (3.8)$$

Similarly, in (3.6) and (3.7), $\lambda_+(t)$ and $\lambda_-(t)$ are initialized as $\lambda_+(t_0) + \lambda_-(t_0) = 1$, $\lambda_+(t_0)\lambda_-(t_0) = 0$ so that $\lambda_+(t) + \lambda_-(t) = 1$ and $\lambda_+(t)\lambda_-(t) = 0$ for any $t \geq t_0$.

Using (2.2), (3.6)–(3.8), we then express the internal signal $z(t)$ as

$$z(t) = \frac{y(t)}{m} + \lambda_+(t)c_+ + \lambda_-(t)c_- + \lambda_0(t)d_0(t) \quad (3.9)$$

where $d_0(t) \in [c_+, c_-]$ is bounded. The term $\lambda_0(t)d_0(t)$ is the part of the signal $z(t)$ which, due to the backlash characteristic, is unobservable from the output $y(t)$. In general, $\lambda_0(t)d_0(t)$ may be nonzero even if the backlash $B(\cdot)$ is known.

Since $z(t)$ is not available, we choose its estimate to be

$$\hat{z}(t) = \frac{y(t)}{m} + \lambda_+(t)c_+ + \lambda_-(t)c_- \quad (3.10)$$

so that the estimation error is $z(t) - \hat{z}(t) = \lambda_0(t)d_0(t)$.

Hence, when the linear part $G(D)$ and the backlash $B(\cdot)$ are known, but the signal $z(t)$ is not available for measurement, the controller (3.5) is modified as

$$u(t) = \theta_u^{*T} \omega_u(t) + \theta_z^{*T} \omega_z(t) + z_m(t + n^*), \quad t \geq t_0 \quad (3.11)$$

where $\omega_z(t) = b(D)[\hat{z}](t)$. Now, despite the estimation error $z(t) - \hat{z}(t) = \lambda_0(t)d_0(t)$, this controller can ensure the exact tracking $e(t) = y(t) - y_m(t) = 0$.

Lemma 3.1 [3]: The controller (3.11) achieves the exact tracking, that is, $y(t + n^*) = y_m(t + n^*)$, for any $t > t_0$, provided that $\hat{z}(\tau) = z(\tau)$ for $\tau = t_0 - n + 1, t_0 - n + 2, \dots, t_0 - 1, t_0, t_0 + 1, \dots, t_0 + n^* - 1, t_0 + n^*$, and $y_m(t_0 + n^*) = y(t_0 + n^*)$.

For the controller (3.11), the initialization for output matching is achieved through the condition: $\hat{z}(\tau) = z(\tau)$ for $\tau = t_0 - n + 1, \dots, t_0 + n^*$, and $y_m(t_0 + n^*) = y(t_0 + n^*)$, where $\hat{z}(t_0 + n^*) = z(t_0 + n^*)$ initializes the backlash inverse mapping: $y(t) \rightarrow \hat{z}(t)$. The condition $\hat{z}(\tau) = z(\tau)$ is satisfied if the motion of $z(\tau)$ does not stay inside the backlash. Since the backlash region characterized by c_+ and c_- is finite, a proper initial excitation of the system can always keep $z(\tau)$ outside of the backlash region.

IV. ADAPTIVE CONTROL FOR UNKNOWN BACKLASH

In this section we propose two adaptive backlash inverse controllers: one for the design with $G(D)$ known, and the other for the design with $G(D)$ unknown. In both designs, the backlash $B(\cdot)$ is unknown and the signal $z(t)$ is not available for measurement.

A. Design for $B(\cdot)$ Unknown and $G(D)$ Known

With $z_m(t)$, $\hat{z}(t)$ defined in (3.3), (3.10), introducing $\theta_m^* = \frac{1}{m}$, $\theta_{c_+}^* = c_+$, $\theta_{c_-}^* = c_-$ and $\omega_y(t) = b(D)[y](t)$, $\omega_+(t) = b(D)[\lambda_+](t)$, $\omega_-(t) = b(D)[\lambda_-](t)$, $\omega_m(t) = \theta_m^{*T} \omega_y(t) + y_m(t + n^*)$, $\omega_{+,m}(t) = \theta_{c_+}^{*T} \omega_+(t) + \lambda_{+,m}(t + n^*)$, $\omega_{-,m}(t) = \theta_{c_-}^{*T} \omega_-(t) + \lambda_{-,m}(t + n^*)$, we express the controller (3.11) as

$$u(t) = \theta_u^{*T} \omega_u(t) + \theta_m^* \omega_m(t) + \theta_{+,m}^* \omega_{+,m}(t) + \theta_{-,m}^* \omega_{-,m}(t). \quad (4.1)$$

We then proceed to use the adaptive version of (4.1) for θ_m^* , $\theta_{c_+}^*$, $\theta_{c_-}^*$ unknown:

$$\hat{u}(t) = \hat{\theta}_u^{*T} \omega_u(t) + \hat{\theta}_m^* \omega_m(t) + \hat{\theta}_{+,m}^* \omega_{+,m}(t) + \hat{\theta}_{-,m}^* \omega_{-,m}(t) \quad (4.2)$$

in which only the estimates $\hat{\theta}_m(t)$, $\hat{\theta}_{c_+}(t)$, $\hat{\theta}_{c_-}(t)$ of θ_m^* , $\theta_{c_+}^*$, $\theta_{c_-}^*$ are updated, because, with $G(D)$ known, we can solve θ_u^* and θ_z^* from (3.4).

To develop an adaptive law for updating the estimates $\hat{\theta}_m(t)$, $\hat{\theta}_{c_+}(t)$, $\hat{\theta}_{c_-}(t)$, dividing both sides of (3.4) by $P(D)$ and operating the resulting identity on $u(t)$, we obtain

$$u(t) = \theta_u^{*T} \omega_u(t) + \theta_z^{*T} \omega_z(t) + z(t + n^*). \quad (4.3)$$

Defining the parameter and regressor vectors as $\theta_b^* = (\theta_m^*, \theta_{c_+}^*, \theta_{c_-}^*)^T$, $\omega_b(t) = (\theta^{*T} \omega_y(t) + y(t + n^*), \theta_{c_+}^{*T} \omega_+(t) + \lambda_{+,m}(t + n^*), \theta_{c_-}^{*T} \omega_-(t) + \lambda_{-,m}(t + n^*))^T$, and using (3.9), we express $u(t)$ in (4.3) as

$$u(t) = \theta_b^{*T} \omega_b(t) + d_1(t + n^*) \quad (4.4)$$

where θ_b^* is the unknown parameter vector, and $d_1(t + n^*) = d_0(t + n^*)\lambda_0(t + n^*) + \theta_z^{*T} b(D)[d_0 \lambda_0](t)$ is the bounded disturbance representing the unobservable part of the backlash input $z(t)$.

With the estimate $\hat{\theta}_b(t)$ of θ_b^* , we use (4.4) to define the estimation error

$$\epsilon_b(t) = \hat{\theta}_b^{*T} \omega_b(t - n^*) + \hat{\theta}_b^T(t - 1) \omega_b(t - n^*) - u(t - n^*).$$

We then choose the adaptive update law for $\hat{\theta}_b(t)$

$$\hat{\theta}_b(t) = \hat{\theta}_b(t - 1) - \frac{\gamma_b \omega_b(t - n^*) \epsilon_b(t)}{1 + \omega_b^T(t - n^*) \omega_b(t - n^*) + \delta^2(t)} + f_b(t) \quad (4.5)$$

where $0 < \gamma_b < 1$, $\delta(t) = (\hat{\theta}_b(t - 1) - \hat{\theta}_b(t - n^*))^T \omega_b(t - n^*)$. To achieve robustness of the parameter adaptation with respect to the bounded disturbance $d_1(t)$, the design signal $f_b(t)$ is to be chosen as one of the existing robustifying modifications [4]–[6]. For example, the switching σ -modification of [5] generates $f_b(t)$ as $f_b(t) = -\sigma(\hat{\theta}_b(t - 1), \sigma_0, M_b) \hat{\theta}_b(t - 1)$, where

$$\sigma(\hat{\theta}_b(t - 1), \sigma_0, M_b) = \begin{cases} \sigma_0 & \text{for } \|\hat{\theta}_b(t - 1)\|_2 > 2M_b \\ 0 & \text{otherwise} \end{cases}$$

with $0 < \sigma_0 < \frac{1-\gamma_b}{2}$. Although not shown in (4.5), with the assumption 5) we use projection to ensure that $\hat{\theta}_m(t) \geq m_0$, $\hat{\theta}_{c_+}(t) \geq 0$ and $\hat{\theta}_{c_-}(t) \leq 0$.

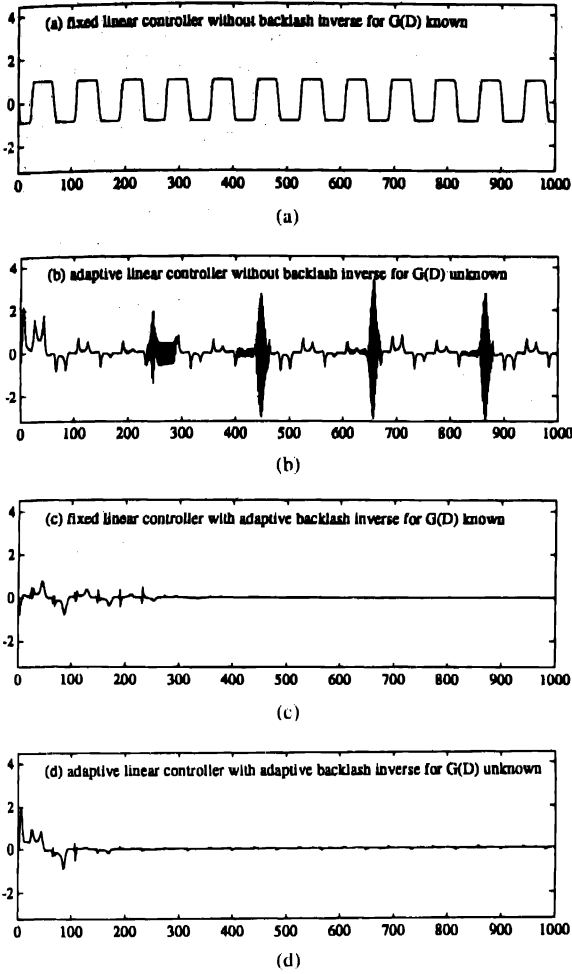


Fig. 3. Tracking errors with different control schemes.

B. Design for $B(\cdot)$ and $G(D)$ Unknown

To develop an adaptive controller structure when both the linear part $G(D)$ and the backlash $B(\cdot)$ are unknown, we define $\theta_y^* = \frac{1}{m}\theta_z^*$, $\theta_r^* = c_r\theta_z^*$, $\theta_l^* = c_l\theta_z^*$, and express (3.11) as

$$u(t) = \theta_u^{*T} \omega_u(t) + \theta_y^{*T} \omega_y(t) + \theta_r^{*T} \omega_r(t) + \theta_l^{*T} \omega_l(t) + \theta_m^* y_m(t+n^*) + \theta_{cr}^* \chi_{rm}(t+n^*) + \theta_{cl}^* \chi_{lm}(t+n^*). \quad (4.6)$$

Since the parameters θ_u^* , θ_y^* , θ_r^* , θ_l^* , θ_m^* , θ_{cr}^* , θ_{cl}^* are unknown, we use their estimates $\theta_u(t)$, $\theta_y(t)$, $\theta_r(t)$, $\theta_l(t)$, $\theta_m(t)$, $\theta_{cr}(t)$, $\theta_{cl}(t)$ to implement the adaptive version of (4.6):

$$u(t) = \theta_u^T(t) \omega_u(t) + \theta_y^T(t) \omega_y(t) + \theta_r^T(t) \omega_r(t) + \theta_l^T(t) \omega_l(t) + \theta_m(t) y_m(t+n^*) + \theta_{cr}(t) \chi_{rm}(t+n^*) + \theta_{cl}(t) \chi_{lm}(t+n^*). \quad (4.7)$$

In this case, we need to update not only the adaptive backlash but also the feedforward and feedback parts of the controller structure (4.7). To develop such an adaptive update law, we introduce $\theta^* = (\theta_u^{*T}, \theta_y^{*T}, \theta_r^{*T}, \theta_l^{*T}, \theta_m^*, \theta_{cr}^*, \theta_{cl}^*)^T$, $\omega(t) = (\omega_u^T(t), \omega_y^T(t), \omega_r^T(t), \omega_l^T(t), y(t+n^*), \chi_{rm}(t+n^*), \chi_{lm}(t+n^*))^T$, and express $u(t)$ in (4.3) as

$$u(t) = \theta^{*T} \omega(t) + d_1(t+n^*). \quad (4.8)$$

With the estimate $\theta(t)$ of θ^* , we use (4.8) to define the estimation error

$$e(t) = \theta^T(t-1) \omega(t-n^*) - u(t-n^*)$$

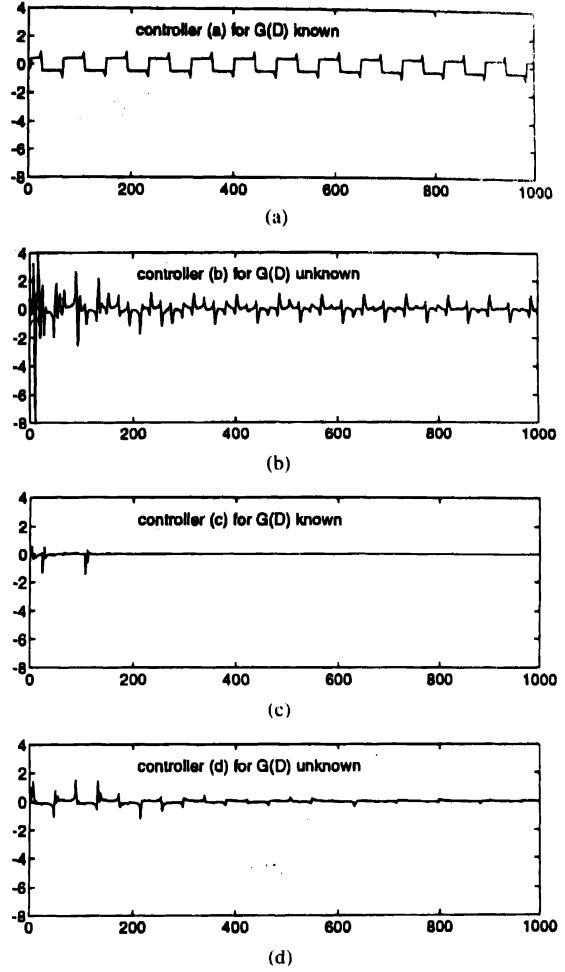


Fig. 4. Tracking errors of the second-order plant.

and use it to construct the update law for $\theta(t)$

$$\theta(t) = \theta(t-1) - \frac{\gamma \omega(t-n^*) e(t)}{1 + \omega^T(t-n^*) \omega(t-n^*)} + f(t) \quad (4.9)$$

where $0 < \gamma < 1$, and $f(t)$ is a design signal for robustness, e.g., $f(t) = -\sigma(\theta(t-1), \sigma_0) M \theta(t-1)$ with $0 < \sigma_0 < \frac{1-\gamma}{2}$. We also use projection to ensure that $\theta_m(t) \geq m_0$, $\theta_{cr}(t) \geq 0$, $\theta_{cl}(t) \leq 0$.

Our adaptive control schemes ensure the bounded-input bounded-output stability, as follows.

Theorem 4.1 [3]: All signals in the closed-loop system consisting of the plant (2.1), the controller (4.2) [or (4.7)], and the update law (4.5) [or (4.9)] are bounded.

Since we have not yet shown that the tracking error $e(t) = y(t) - y_m(t)$ converges to zero, the tracking performance of the adaptive systems will be illustrated by simulation results.

V. EXAMPLES

We now use examples to illustrate our adaptive designs and present simulation results which show that our adaptive control schemes lead to significant improvements of tracking performance.

Example 5.1: We consider $G(D) = \frac{1}{D+a_1}$ where $a_1 = 1.83$. The backlash $B(\cdot)$ with $m = 0.014$, $c_r = 22.5$, and $c_l = -24.7$ is unknown to the following controllers.

a) Fixed linear controller with no backlash inverse for $G(D)$ known:

$$u(t) = \theta_m^* \theta_z^* y(t) + \theta_m^* y_m(t+1), \quad \theta_m^* = \frac{500}{7}.$$

b) Adaptive linear controller with no backlash inverse for $G(D)$ unknown:

$$u(t) = \theta_y(t)y(t) + \theta_m(t)y_m(t+1).$$

c) Fixed linear controller with an adaptive backlash inverse (see Section IV-A):

$$u(t) = \theta_m(t)(\theta_y^* y(t) + y_m(t+1)) + \theta_{i,1}(t)(\theta_{i,1}^* \lambda_1(t) + \lambda_{1,m}(t+1)) + \theta_{i,2}(t)(\theta_{i,2}^* \lambda_2(t) + \lambda_{2,m}(t+1)).$$

d) Adaptive linear controller with an adaptive backlash inverse (see Section IV-B):

$$u(t) = \theta_y(t)y(t) + \theta_{i,1}(t)\lambda_1(t) + \theta_{i,2}(t)\lambda_2(t) + \theta_m(t)y_m(t+1) + \theta_{i,1}(t)\lambda_{1,m}(t+1) + \theta_{i,2}(t)\lambda_{2,m}(t+1).$$

The initial parameter estimates are $\theta_y(0) = \theta(0)\theta_m(0)$, $\theta_{i,1}(0) = \theta_{i,1}(0)$, $\theta_{i,2}(0) = \theta_{i,2}(0)$, $\theta(0) = 2.5$, $(\theta_m(0), \theta_{i,1}(0), \theta_{i,2}(0))^T = (70, 10, -10)^T$. Their matching values are $\theta_y^* = \theta^*\theta_m^*$, $\theta_{i,1}^* = \theta^*\theta_{i,1}^*$, $\theta_{i,2}^* = \theta^*\theta_{i,2}^*$, $\theta^* = 1.83$, $(\theta_m^*, \theta_{i,1}^*, \theta_{i,2}^*)^T = (\frac{500}{7}, 22.5, -24.7)^T$.

Example 5.2 We now consider the second-order $G(D) = \frac{1}{D^2 + a_2 D + a_1}$ with $a_2 = -1.3$, $a_1 = -0.3$. The backlash $B(\cdot)$ is the same as in Simulation 1. For this $G(D)$, we have the following.

Controller a):

$$u(t) = \theta_u^* u(t-1) + \theta_m^* \theta_{i,1}^{*-1} (y(t-1), y(t))^T + \theta_m^* y_m(t+2).$$

Controller b):

$$u(t) = \theta_u(t)u(t-1) + \theta_y^T(t)(y(t-1), y(t))^T + \theta_m(t)y_m(t+2).$$

Controller c):

$$u(t) = \theta_u^* u(t-1) + \theta_m(t)(\theta^{*-1} (y(t-1), y(t))^T + y_m(t+2)) + \theta_{i,1}(t)(\theta^{*-1} (\lambda_1(t-1), \lambda_1(t))^T + \lambda_{1,m}(t+2)) + \theta_{i,2}(t)(\theta^{*-1} (\lambda_2(t-1), \lambda_2(t))^T + \lambda_{2,m}(t+2)).$$

Controller d):

$$u(t) = \theta_u(t)u(t-1) + \theta_y^T(t)(y(t-1), y(t))^T + \theta_{i,1}^T(t)(\lambda_1(t-1), \lambda_1(t))^T + \theta_{i,2}^T(t)(\lambda_2(t-1), \lambda_2(t))^T + \theta_m(t)y_m(t+2) + \theta_{i,1}(t)\lambda_{1,m}(t+2) + \theta_{i,2}(t)\lambda_{2,m}(t+2).$$

The initial parameters are: $(\theta_u(0), \theta_y^T(0), \theta_m(0)) = (\theta_u^*, \theta^{*-1} \theta_m^*, \theta_m^*)$ for b); $(\theta_m(0), \theta_{i,1}(0), \theta_{i,2}(0)) = (80.5, 21, -21)$ for c); and $\theta_u(0) = -1.2$, $\theta_y(0) = 80.5\theta(0)$, $\theta_{i,1}(0) = 24.9\theta(0)$, $\theta_{i,2}(0) = -25.3\theta(0)$, $\theta(0) = (-0.3, -1.5)$, $(\theta_m(0), \theta_{i,1}(0), \theta_{i,2}(0)) = (80.5, 24.9, -25.3)$ for d). Their matching values are $\theta_u^* = -1.3$, $\theta^* = (-0.39, -1.99)^T$, $(\theta_m^*, \theta_{i,1}^*, \theta_{i,2}^*)^T = (\frac{500}{7}, 22.5, -24.7)^T$.

Typical responses of these control schemes to $y_m(t) = 15 \sin(0.0754t)$, $\gamma = 0.1$ and $\sigma_0 = 0.02$ are given in Fig. 3 for Example 5.1, and in Fig. 4 for Example 5.2. The simulation results show that the control schemes a) and b) which ignore backlash lead to large tracking errors $e(t) = y(t) - y_m(t)$. Our adaptive backlash inverse control schemes c) and d), which take into account the effects of the unknown backlash, lead to very small tracking errors.

VI. CONCLUSIONS

In this note we have developed two adaptive control schemes for systems with a backlash characteristic at the output. The first scheme is for systems with a known linear part and an unknown backlash. The second scheme is for a plant with both the linear part and backlash unknown. Our adaptive controller structures consist of a linear feedforward part and a linear-like feedback part which incorporates an adaptive backlash inverse. These controller structures result in linear parameterizations from which adaptive laws can be designed to update the controller parameters to ensure the closed-loop signal boundedness. Simulations results show that significantly improved system performance can be achieved by our adaptive control schemes.

REFERENCES

- [1] G. Tao and P. V. Kokotović, "Adaptive control of systems with backlash," *Automatica*, vol. 29, no. 2, pp. 323-335, 1993.
- [2] —, "Continuous-time adaptive control of systems with unknown backlash," in *Proc. 1993 Amer. Contr. Conf.* San Francisco, CA, pp. 1344-1348.
- [3] —, "Adaptive adaptive control of systems with unknown output backlash," Tech. Rep. CCEC-921201, UCSB, Dec. 1992.
- [4] G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
- [5] P. A. Ioannou and K. Tsakalis, "Robust discrete time adaptive control," in *Adaptive and Learning Systems: Theory and Applications*. K. S. Narendra, Ed. New York: Plenum, 1986.
- [6] G. Kreisselmeier and B. D. O. Anderson, "Robust model reference adaptive control," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 127-133, 1986.

Reciprocal Processes on a Tree—Modeling and Estimation Issues

Robert W. Dijkerman, Ravi R. Mazumdar, and Arunabha Bagchi

Abstract—Motivated by multiresolution decomposition methods such as the discrete wavelet transformation, we introduce reciprocal processes on truncated N -ary trees. We discuss the relationship between such processes and nearest neighbor models. We show that we can derive a recursive description of the process, and that all reciprocal processes on N -ary trees reduce to autoregressive processes in the case of zero-valued boundary values at the bottom of the tree, corresponding to truncation of the tree. We then study the smoothing equations associated with such models.

Manuscript received December 1, 1993; revised May 23, 1994. This work was supported in part by a grant from Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (FCAR). Part of this work was presented at the 7th SP Workshop on Statistical Signal and Array Processing, Québec City, Canada, June 26-29, 1994.

R. W. Dijkerman was with the INRS-Télécommunications, Université du Québec, Ile des Soeurs, Verdun, P.Q., H3E 1H6, Canada and is now with the National Aerospace Laboratory NLR, PO Box 90502, 1006 BM Amsterdam, The Netherlands.

R. R. Mazumdar is with the INRS-Télécommunications, Université du Québec, Ile des Soeurs, Verdun, P.Q., H3E 1H6, Canada.

A. Bagchi is with the University of Twente, Faculty of Applied Mathematics, Enschede, The Netherlands.

IEEE Log Number 9406997

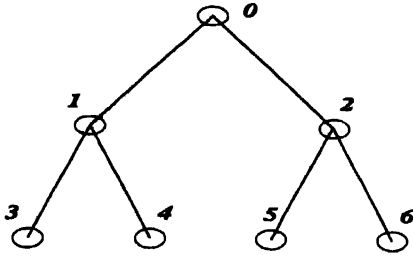


Fig. 1 Simple truncated binary tree (with numbering of nodes)

1 INTRODUCTION

Many multiresolution signals can be described on truncated λ -ary trees. For instance, the discrete wavelet coefficients of an n -dimensional signal can be identified with the nodes on a truncated 2-ary tree. We show an example of such an λ -ary tree in Fig. 1 for $\lambda = 2$. Each node on the tree, except for the sets of initial nodes in the top and the final nodes at the bottom of the tree, has one 'parent' node and λ 'offspring' nodes. Note that the simplest tree, where $\lambda = 1$, can be identified with an interval $[Z_1, Z_2]$ on the set of integers.

In [4], Dijkerman and Mazumdar showed that for a wide class of second-order processes, there is nonzero correlation between wavelet coefficients at all scales with strong decay in the correlation when coefficients are located further away from each other in the wavelet domain. We are therefore interested in the class of stochastic processes on truncated λ -ary trees, having a so-called reciprocal property. That is, the conditional probability of the value of a node on the tree, given the values of all other nodes in the tree, is only dependent on the values of the parent node and the offspring nodes. Basseville *et al.* [1] introduced autoregressive processes on λ -ary trees, where a node is conditionally dependent only on its parent node. Although causality is very well motivated in time, this is not really the case in the time-frequency domain. The conditional dependence of a time-frequency coefficient on the scale above it is not obvious.

Discrete time reciprocal processes on an interval were studied in detail by Levy *et al.* [8]. In this note, we closely follow their approach and extend their results for the case $\lambda > 1$. We like to mention as well the work by Greenc and Levy [5], who studied in detail several different solution procedures of the smoother problem for discrete time reciprocal processes, and the work by Levy, Fabre, and Benveniste [7], who studied independently a class of Markov random fields on trees. Their tree structure is slightly more complicated, but techniques and derivations similar to the work in this note are reported. They study a smoothing problem and discuss several iterative solution techniques. We refer to [3] for a more detailed version of this note.

II RECIPROCAL PROCESSES AND NEAREST NEIGHBOR MODELS ON TREES

In this section, we define the notion of reciprocity of processes on trees and establish the relation between such processes and so-called nearest neighbor models.

Let us denote by $t, t_{-1}, \dots, t_{-\lambda}$, respectively, the parent and the λ offspring nodes of the node t in the truncated tree T , where t does not belong to I , the set of root nodes, and I_f , the set of final nodes of T . For simplicity of notation, we now consider processes on binary trees only, although the case where $\lambda > 2$ is a straightforward extension of $\lambda = 2$. We shall say that a stochastic process $S(\cdot)$, defined on a truncated binary tree T , taking values in R^n , is reciprocal

if and only if

$$P[S(t) \leq s(t) | S(s) = s(s), \quad s \in T \setminus \{t\}] \\ = P[S(t) \leq s(t) | S(t) = s(t), S(t_{-1}) = s(t_{-1}), \dots, S(t_{-\lambda}) = s(t_{-\lambda})] \quad (1)$$

for $t \in \text{int}(I)$ (i.e., $t \in I, t \neq \{T, T_f\}$), where $P(\cdot)$ denotes the probability.

Assume we have a zero mean Gaussian reciprocal process $S(\cdot)$. We project $S(t), t \in \text{int}(T)$, on all other stochastic variables on the tree:

$$L[S(t) | S(s), \quad s \in T \setminus \{t\}] \\ = P_{-1}(t)S(t) + P_1(t)S(t_{-1}) + P_{-2}(t)S(t_{-2}) + \dots \quad (2)$$

The projection implies that the so-called residual process

$$d(t) = S(t) - (P_{-1}(t)S(t) + P_1(t)S(t_{-1}) + P_{-2}(t)S(t_{-2}) + \dots) \quad (3)$$

is orthogonal to $S(s), t \neq s$.

The relation (3) specifies a Nearest Neighbor Model (NNM) for $S(\cdot)$, where the driving noise is the residual process $d(\cdot)$, having the following correlation structure:

$$\begin{cases} D(t) - \Gamma[d(t)d(t)^T] = \sigma_{d(t)} \\ \Gamma[d(t)d(t_{-1})^T] = -\sigma_{d(t)}^T P_{-1}^T(t) - P_{-1}(t)\sigma_{d(t_{-1})} \\ L[d(t)d(s)^T] = 0, \quad s \neq t, t_{-1}, \dots, t_{-\lambda} \end{cases} \quad (4)$$

where $\sigma_{d(t)}$ is the covariance matrix of the random vector $d(t)$. The correlation structure of $d(t)$ tells us that it is a moving average process (defined on T). We shall discuss this later in detail. Property (4) shows that the projection matrices P_{-1}, \dots , and the noise variance $D(\cdot)$ cannot be specified independently of each other. Standard Gaussian estimation theory implies that the existence and unicity of P_{-1}, \dots is guaranteed if the covariance matrix of the vector $[S(t)S(t_{-1})S(t_{-2})]$ is positive definite for all $t \in \text{int}(T)$. The model described by (3) specifies the process $S(\cdot)$ on $\text{int}(I)$. The residual process $d(t), t \in \text{int}(I)$, is uncorrelated with all $S(s), s \in \{T, I_f\}$. Therefore, we choose independently of the residual process, a vector b of Dirichlet boundary values (corresponding to all boundary nodes) having an arbitrary covariance matrix, which we shall denote B .

The truncation of a tree implicitly implies that the values of the nodes lower in the tree are zero or independent of the truncated tree. Therefore, as an extra assumption, we could demand the following autoregressive property to hold for $t \in I_f$:

$$d(t) = S(t) - P_{-1}(t)S(t_{-1}) \quad (5)$$

with $d(t)$ orthogonal to $S(s)$ for $t \neq s$. That is, the final nodes have an autoregressive property. This is equivalent to the assumption of a reciprocal process on an extended tree T having a layer of zero-valued final nodes.

It can be verified following [8] that the model (3) with the specified noise structure and the Dirichlet boundary conditions is well defined. There is a unique solution if the covariance matrix R of the vector of interior nodes values is positive definite.

We show now that an NNM, as in (3), with noise structure as in (4) and the Dirichlet boundary conditions implies necessary reciprocity. Let us assume we have the following Nearest Neighbor Model for $t \in \text{int}(T)$:

$$M_0(t)S(t) - M_{-1}(t)S(t_{-1}) - M_1(t)S(t_{-2}) - \dots = \epsilon(t) \quad (6)$$

such that $M_{-i}(t)^T = M_i(t)$ for $i = 1, 2, \dots$, being square matrices of size n . We assume a Gaussian distributed vector of root and final nodes values b having covariance matrix B . The input noise $\epsilon(t)$ is

a Gaussian process uncorrelated with the boundary process b and has the following correlation structure:

$$\begin{cases} E[e(t)e(t)'] = M_0(t) \\ E[e(t)e(t_{\gamma_1})'] = -M_1(t) \\ E[e(t)e(s)'] = 0 \quad s \neq t, \hat{t}, t_{\gamma_1}, t_{\gamma_2}. \end{cases} \quad (7)$$

Furthermore, it is assumed that the system is well posed, i.e., it admits a unique solution. The relation between the original NNM and this model is easily established. Assume that $S(\cdot)$ is a reciprocal process with driving noise process $d(\cdot)$, such that $D(t)$ is invertible for all $t \in \text{int}(T)$. The multiplication of (3) by $D(t)^{-1}$ and the definitions $M_0(t) = D(t)^{-1}$, $M_-(t) = D(t)^{-1}P_-(t)$, $M_i(t) = D(t)^{-1}P_i(t)$, and $e(t) = D(t)^{-1}d(t)$ results in (6). We can now easily verify that $M_-(t_{\gamma_1})' = M_1(t)$ and that (7) holds.

This shows that any reciprocal process $S(\cdot)$ on T with $R > 0$ gives rise to an NNM of the form (6). Now we have to show that a well-posed NNM of the form (6), (7) is necessarily reciprocal. Let us therefore define the difference operator associated to (6) by

$$\Lambda = M_0(t)I - M_-(t)Z^{-1} - M_1(t)Z^1 - M_2(t)Z^2 \quad (8)$$

where Z^{-1} denotes the backward shift on the binary tree $Z^{-1}f(t) = f(\hat{t})$, and Z^i , $i = 1, 2$ denotes the forward shift $Z^i f(t) = f(t_{\gamma_i})$.

The operator Λ maps n -vector functions defined on T into n -vector functions defined on the interior of T . Consider now the L^2 -space of n -vector functions defined over the interior of T , with the inner product $\langle f, g \rangle = \sum_{s \in \text{int}(T)} f(s)' g(s)$. In Appendix A, we show that the following Green's identity holds:

$$\begin{aligned} \langle f, \Lambda S(\cdot) \rangle - \langle \Lambda f, S(\cdot) \rangle \\ = \sum_{i \in I_i} [f(t)' f(t_{\gamma_1})' f(t_{\gamma_2})'] E_i^j \begin{bmatrix} S(t) \\ S(t_{\gamma_1}) \\ S(t_{\gamma_2}) \end{bmatrix} \\ + \sum_{i \in I_j} [f(t)' f(\hat{t})'] E_i^j \begin{bmatrix} S(t) \\ S(t) \end{bmatrix} \end{aligned} \quad (9)$$

where the matrices E_i^j and E_i^j are defined as follows:

$$E_i^j = \begin{bmatrix} 0 & M_1(t) & M_2(t) \\ -M_-(t_{\gamma_1}) & 0 & 0 \\ -M_-(t_{\gamma_2}) & 0 & 0 \end{bmatrix} \quad (10)$$

and

$$E_i^j = \begin{bmatrix} 0 & M_-(t) \\ -M_-(t)' & 0 \end{bmatrix}. \quad (11)$$

The Green's function Γ is defined by

$$\begin{cases} \Lambda \Gamma(t, s) = I \delta(t, s) & t, s \in \text{int}(T) \\ \Gamma(t, s) = 0 & t \text{ or } s \in \{T_i, T_j\}. \end{cases} \quad (12)$$

We show also in Appendix A that $\Gamma(t, s) = \Gamma(s, t)'$ for $t, s \in T$.

Following exactly the work of Levy *et al.* [8], we derive for the N -ary tree that

$$\begin{aligned} S(k) = \sum_{s \in \text{int}(T)} \Gamma(k, s) e(s) + \sum_{i \in I_i} (\Gamma(k, t_{\gamma_1}) M_-(t_{\gamma_1}) \\ + \Gamma(k, t_{\gamma_2}) M_-(t_{\gamma_2})) S(t) + \sum_{i \in I_j} \Gamma(k, \hat{t}) M_-(t)' S(t). \end{aligned} \quad (13)$$

Equation (13) can be used to show that the solution $S(k)$ is orthogonal to $e(s)$ for $k \neq s$

$$\begin{aligned} E[S(k)e(s)'] &= \sum_{i \in \text{int}(T)} \Gamma(k, i) E[e(i)e(s)'] = (\Lambda \Gamma(s, k))' \\ &= I \delta(s, k). \end{aligned} \quad (14)$$

Now we can state and prove the following.

Theorem 2.1: Let $S(\cdot)$ be a zero-mean Gaussian process on a binary tree, having covariance matrix $R > 0$. Then, $S(\cdot)$ is reciprocal if and only if it admits a well-posed NNM of the form (6), (7).

Proof of Theorem 2.1: We have shown in this section that if $S(\cdot)$ is reciprocal, we can write the process in the form (6), (7). Now we must show that the solution $S(k)$ of (6), (7), given by (13), is reciprocal. To do this, let T' be a subtree of T (having the same structure as T), and let l and k be two nodes of T , such that $k \in \text{int}(T')$ and $l \notin T'$.

To prove that $S(\cdot)$ is reciprocal, we have to show that $\hat{S}(k) = S(k) - E[S(k)|S(t), t \in \{T'_i, T'_j\}] \perp S(l)$. The orthogonality principle derived above says that for $s \in \text{int}(T')$ and for $t \in \{T'_i, T'_j\}$, $e(s) \perp S(t)$. Therefore, using the solution of (6) on the subtree T' [similar to (13)], we obtain

$$\begin{aligned} E[S(k)|S(t), t \in \{T'_i, T'_j\}] &= \sum_{i \in I'_i} (\Gamma(k, t_{\gamma_1}; T') M_-(t_{\gamma_1}) \\ &+ \Gamma(k, t_{\gamma_2}; T') M_-(t_{\gamma_2})) S(t) + \sum_{i \in I'_j} \Gamma(i, k; T')' M_-(t)' S(t) \end{aligned} \quad (15)$$

where $\Gamma(k, s; T')$ is the Green's function for the model on the subtree T' . So $\hat{S}(k) = \sum_{s \in \text{int}(T')} \Gamma(k, s; T') e(s)$, which shows that $\hat{S}(k) \perp S(l)$, and therefore that the process is reciprocal. This completes the proof.

III. RECURSIVE DESCRIPTION OF A RECIPROCAL PROCESS

The NNM described in (6), (7), or (13) is nonrecursive. We derive in this section a recursive description.

We construct a moving average representation of the driving noise $e(t)$. For this, we introduce $w(t)$, a white Gaussian noise (WGN) process with intensity $W(t)$, which is uncorrelated with the boundary process b of the reciprocal process. Let us define $e(t)$ by introducing the matrix $A(t)$ in the following way:

$$e(t) = W(t)^{-1} w(t) - A(t_{\gamma_1})' W(t_{\gamma_1})^{-1} u(t_{\gamma_1}) - A(t_{\gamma_2})' W(t_{\gamma_2})^{-1} u(t_{\gamma_2}) \quad (16)$$

for $t \in \text{int}(T)$. Then, the relations (7) for the covariance of $e(t)$ imply that the matrices $A(t)$ and $W(t)$ must satisfy, for $t \in \text{int}(T)$,

$$\begin{aligned} M_0(t) &= W(t)^{-1} + A(t_{\gamma_1})' W(t_{\gamma_1})^{-1} A(t_{\gamma_1}) \\ &+ A(t_{\gamma_2})' W(t_{\gamma_2})^{-1} A(t_{\gamma_2}) \end{aligned} \quad (17)$$

and for $i = 1, 2$, $t, t_{\gamma_i} \in \text{int}(T)$

$$M_i(t) = -E[e(t)e(t_{\gamma_i})'] = A(t_{\gamma_i})' W(t_{\gamma_i})^{-1}. \quad (18)$$

We can now substitute, for $t, t_{\gamma_i} \in \text{int}(T)$, $A(t_{\gamma_i}) = W(t_{\gamma_i}) M_i(t)'$. We define, by choice of an arbitrary symmetric positive definite matrix $W(t_{\gamma_i})$, $A(t_{\gamma_i}) = W(t_{\gamma_i}) M_i(t)'$ for $t_{\gamma_i} \in T_j$. We see that $W(t)$ must satisfy, for $t \in \text{int}(T)$, the backward difference equation

$$W(t)^{-1} = M_0(t) - M_1(t) W(t_{\gamma_1}) M_1(t)' - M_2(t) W(t_{\gamma_2}) M_2(t)'. \quad (19)$$

Since we want the solution $W(t)$ to be a covariance matrix for each $t \in \text{int}(T)$, it must be a nonnegative matrix. We show in Appendix B, that this indeed holds.

The noise representation in (16) satisfies (7) if we use (19). The NNM (6), (7) also implies that, for $t \in \text{int}(T)$,

$$M_-(t) = (A(t)' W(t)^{-1})' = W(t)^{-1} A(t). \quad (20)$$

Consider now the operators

$$\begin{cases} \Omega = I - A(t)Z^{-1} \\ W = W(t)I \\ \Omega^* = I - A(t\gamma_1)^T Z^{-1} - A(t\gamma_2)^T Z^T \end{cases} \quad (21)$$

then we derive, for $t \in \text{int}(T)$

$$\begin{cases} \Omega^* W^{-1} u(t) = \Omega^* W(t)^{-1} u(t) = \epsilon(t) \\ \Omega^* W^{-1} \Omega S(t) = M_0(t)S(t) - M_1(t)S(t\gamma_1) \\ \quad - M_2(t)S(t\gamma_2) - W(t)S(t) \end{cases} \quad (22)$$

and thus the relation $\Lambda S(t) = \epsilon(t)$ gives us $\Omega^* W^{-1}(\Omega S(t) - \epsilon(t)) = 0$. Define now for $t \in T \setminus I$

$$\begin{aligned} Z(t) &= W^{-1}(\Omega S(t) - u(t)) \\ &= W(t)^{-1}(S(t) - A(t)S(t) - u(t)) \end{aligned} \quad (23)$$

Then we know that $\Omega^* Z(t) = 0$ $t \in \text{int}(T)$. This together with (23) results in the following

Proposition 3.1 The set of equations

$$\begin{cases} Z(t) = A(t\gamma_1)^T Z(t\gamma_1) + A(t\gamma_2)^T Z(t\gamma_2) & \text{for } t \in \text{int}(T) \\ S(t) = A(t)S(t) + u(t) + W(t)Z(t) & \text{for } t \in T \setminus I \end{cases} \quad (24)$$

together with the Dirichlet boundary conditions introduced before describe the NNM (6), (7).

Let us assume we have the solution of (24) with the boundary conditions $Z_0(t) = 0$ $t \in I_f$ and $S_0(t) = b(t)$ $t \in I$. We get $Z_t(t) = 0$ for all t of interest and thus $S_0(t)$ is an autoregressive process (on a tree) which can be computed recursively by performing the last part of (24). This process satisfies (6) and the correct initial condition but its end values are generally different from b_f the final node values. If $\Gamma(t, s)$ is the Green's function the solution of (6) (7) for $t \in \text{int}(T)$ is therefore given by

$$S(t) = S_0(t) + \sum_{s \in I_f} \Gamma(t, s) M(s)^T (S(s) - S_0(s)) \quad (25)$$

The correction formula requires the precomputation of $\Gamma(t, s)$ $t \in T$ $s \in I_f$.

Let us now discuss the assumption of the autoregressive property in the final nodes as in (5). This assumption is equivalent to the assumption of an extended tree \bar{T} with zero valued final nodes. Computing for the extended tree \bar{T} the variable $Z(t)$ $t \in \bar{T}_f$ (or $t \in T \setminus I_f$) using (16) (20) and (23) gives us

$$\begin{aligned} Z(t) &= W(t)^{-1}(S(t) - A(t)S(t) - u(t)) \\ &= M_0(t)S(t) - M_1(t)S(t\gamma_1) - \epsilon(t) = 0 \end{aligned} \quad (26)$$

resulting in using (24) $Z(t) = Z_0(t) = 0$ $t \in \bar{T} \setminus \bar{T}_f$. Finally this results in $S(s) = S_0(s)$ for $t \in \bar{T}$. That is if we assume we have the autoregressive property in the final nodes of the original tree, we can rewrite the reciprocal process as an autoregressive process on the tree for all $t \in \bar{T}$. Of course the reverse is true as well leading to the following

Proposition 3.2 A reciprocal process on the tree is an autoregressive process on the entire tree if and only if the reciprocal process has the autoregressive property in the final nodes

IV A SMOOTHING PROBLEM

We shall now apply the previous definitions and insights to the estimation problem. Consider a reciprocal process $S(\cdot)$ defined on T , having zero-valued final nodes, described by the model (6) (7). We are given the observations $Y(t) = H(t)S(t) + \epsilon(t)$ where $\epsilon(\cdot)$ is a WGN process, uncorrelated with $S(\cdot)$ having intensity $\Lambda(\cdot)$. We want to compute the smoothed estimate in all nodes of the tree $\hat{S}(t) = E[S(t)|Y]$, where Y is the vector of all observations. We derive, for $t \in \text{int}(T)$, following the derivation of Levy [6] using iterated conditioning

$$\begin{aligned} M_0(t)S(t) &= M_-(t)S(t) + M_1(t)S(t\gamma_1) \\ &\quad + M_2(t)S(t\gamma_2) + H(t)^T \Lambda(t)^{-1} (Y(t) - H(t)S(t)) \end{aligned} \quad (27)$$

We have a difference equation for the smoothed estimates which we shall solve recursively. Equation (27) implies that we have the following model for the smoothed estimates in the interior of T

$$\Lambda_I S(t) = H(t)^T \Lambda(t)^{-1} Y(t) \quad (28)$$

with

$$\begin{cases} \Lambda_I = M_{0I}(t)I - M_1(t)Z^{-1} - M_2(t)Z^{-1} - M_1(t)Z^{-1} \\ M_{0I}(t) = M_0(t) + H(t)^T \Lambda(t)^{-1} H(t) \end{cases} \quad (29)$$

Then, following the procedure of the previous section the operator Λ_I can be represented as $\Lambda_I = \Omega_I^T W_I^{-1} \Omega_I$ with

$$\begin{cases} \Omega_I = I - A_I(t)Z^{-1} \\ W_I = W_I(t)I \\ \Omega_I^* = I - A_I(t\gamma_1)^T Z^{-1} - A_I(t\gamma_2)^T Z^T \end{cases} \quad (30)$$

where $A_I(t) = W_I(t\gamma_1)M(t)^T$ for $t \in \text{int}(T)$ ($W_I(t)$ is an arbitrary symmetric positive definite matrix for $t \in I_f$) and

$$\begin{aligned} W_I(t)^{-1} &= M_{0I}(t) - M_1(t)W_I(t\gamma_1)^{-1}M_1(t)^T \\ &\quad - M_2(t)W_I(t\gamma_2)^{-1}M_2(t)^T \end{aligned} \quad (31)$$

for $t \in \text{int}(I)$. Define now for $t \in T \setminus I$

$$Z_I(t) = W_I(t)^{-1} \Omega_I S(t) = W_I(t)^{-1} (S(t) - A_I(t)S(t)) \quad (32)$$

Then (28) leads to $\Omega_I^T Z_I(t) = \Lambda(t)^{-1} Y(t)$ for $t \in \text{int}(T)$.

Proposition 4.1 The smoother equations (28) can be rewritten as (33), found at the bottom of the page. These two equations propagate causally in the backward and forward directions (on the tree) respectively, describing the structure of the smoother.

Now let $Z_{0I}(t)$ and $S_0(t)$ be the solution of (33) with boundary conditions $Z_{0I}(s) = 0$ $s \in T_f$ and $S_0(s) = Y(s)$ $s \in T$. The smoothed estimate we obtain does not take into account the boundary conditions in \bar{T} and T_f . Following the solution (13) we can conclude that the correction formula for the boundary conditions then becomes

$$\begin{aligned} S(t) &= S_0(t) + \sum_{s \in I_f} (\Gamma_I(t, s\gamma_1)M(s\gamma_1) \\ &\quad + \Gamma_I(t, s\gamma_2)M(s\gamma_2))(S(s) - Y(s)) \\ &\quad + \sum_{s \in I_f} \Gamma_I(t, s)M(s)^T (S(s) - S_0(s)) \end{aligned} \quad (34)$$

where Γ_I is the Green's function for (28)

$$\begin{cases} Z_I(t) = A_I(t\gamma_1)^T Z_I(t\gamma_1) + A_I(t\gamma_2)^T Z_I(t\gamma_2) + \Lambda(t)^{-1} Y(t) \\ \quad \text{for } t \in \text{int}(T) \\ S(t) = A_I(t)S(t) + W_I(t)Z_I(t) \\ \quad \text{for } t \in T \setminus T_f \end{cases} \quad (33)$$

If we are able to observe $S(s)$ exactly at the boundary nodes, we have

$$\hat{S}(t) = \hat{S}_0(t) + \sum_{s \in T_f} \Gamma_E(t, \hat{s}) M_{-}(s)^T (\hat{S}(s) - \hat{S}_0(s)). \quad (35)$$

In the case of an autoregressive process, or an extended tree \bar{T} with zero-valued final nodes, we have $A_{T_f}(t\gamma_i) = W_{E_f}(t\gamma_i) M_i(t)^T = 0$ for $t\gamma_i \in \bar{T}_f$ (since $M_i(t) = 0$). Therefore, since we have chosen $Z_{0E}(s) = 0$ for $s \in \bar{T}_f$, we get $\hat{S}(s) = \hat{S}_0(s) = 0$ for $s \in \bar{T}_f$. Then, if we are able to observe $S(s)$ exactly at the boundary nodes in T_f , we have $\hat{S}(t) = S_0(t)$, $t \in T$ and we do not need a correction equation as in (34) or (35). The smoothing reduces in this case to a simple double-sweep algorithm.

The Rauch-Tung-Striebel (RTS) algorithm for autoregressive processes on trees, developed by Chou *et al.* [2], is quite similar to the double-sweep algorithm here. The difference with Chou's algorithm is that the true value of $\hat{S}(s)$ for $s \in T_f$ is obtained in their upward sweep (T_f should, however, consist of just one node), and there is no need for a correction equation. We need to observe exactly $S(s)$ for $s \in T_f$, or apply the correction equation (34).

Let us study the smoothing error $\hat{S}(t) = S(t) - \hat{S}(t)$. We derive, using (6) and (28), that $\Lambda_E \hat{S}(t) = \Lambda_E S(t) - \Lambda_F \hat{S}(t) = \hat{e}(t)$, with $\hat{e}(t)$ having a covariance structure similar to $e(t)$, due to the orthogonality of e and v . Consider now the boundary process $\hat{S}(t) = S(t) - \hat{S}(t) = Y(t) - v(t) - \hat{S}(t)$ for $t \in T_f$. Linear estimation theory tells us that $\hat{S}(t)$ is orthogonal to all $Y(s)$, $s \in T$. Therefore, the driving noise $\hat{e}(s) = \Lambda_E \hat{S}(s)$ for $s \in \text{int}(T)$ is orthogonal to all $Y(s)$, $s \in T$. The smoothing error $\hat{S}(t)$, $t \in T_f$, can be written as a linear combination of $Y(s)$, $s \in T$ and $v(t)$, and is therefore orthogonal to $\hat{e}(s)$, $s \in \text{int}(T)$. Thus, we have proved the following.

Proposition 4.2: The smoothing error $\hat{S}(t) = S(t) - \hat{S}(t)$ of a reciprocal process on a tree is itself reciprocal.

Assume now that we have an autoregressive process, described as a reciprocal process on an extended tree, with zero-valued final nodes. Therefore, the error of the smoothing problem on the extended tree (without necessarily observing $S(t)$ in the final and root nodes of the original tree) is again a reciprocal process on an extended tree with zero-valued final nodes. We can conclude the following.

Corollary 4.1: The smoothing error of a first-order autoregressive process on a tree is itself a first-order autoregressive process.

Remark: Luetgen [9] proved this result directly for autoregressive processes on a tree, obtaining explicitly the model parameters. We derived the result as a byproduct of our study of reciprocal processes.

V. CONCLUSION

Motivated by multiresolution decompositions of signals, we have introduced reciprocal processes on truncated N -ary trees; and following the results of Levy, Frezza, and Krener [8], we have described nearest neighbor models and recursive structures for these processes. We found that reciprocal processes with zero-valued final nodes reduce to autoregressive processes on a tree, which motivates, for

another time, the study of these processes in, for example, [1] and [2]. We described the smoother equations for such processes.

Reciprocal processes could model in an efficient way many interesting stochastic phenomena on multiple scales. Their recursive structure makes them suitable for fast signal processing algorithms of processes. The more general study of Markov random field processes on trees (having arbitrary neighborhood structures, including possible neighbors on the same scale in the tree) is now being pursued.

APPENDIX A

In this appendix, we prove the Green's identity used in Section II, and show that the Green's function $\Gamma(t, s)$ is symmetric, i.e., $\Gamma(t, s) = \Gamma(s, t)^T$, for $t, s \in T$.

We prove the Green's identity by means of an example which can be generalized for any reciprocal process on a truncated N -ary tree. Let us consider the very simple binary tree as in Fig. 1, using the indicated numbering of the nodes. We define the matrix $\hat{\Lambda}$ as shown in (A.1), found at the bottom of the page.

Using $M_{-}(t\gamma_i)^T = M_i(t)$, we see that $\hat{\Lambda}$ is a symmetric matrix. We also see that the rows of this matrix relating to the interior nodes of T describe the operator Λ in matrix form. Using the symmetry of $\hat{\Lambda}$, we have for any appropriate dimensional vector f (in this case, 7-dimensional) that the equality $f^T \hat{\Lambda} S = (\hat{\Lambda} f)^T S$ holds. Now we derive, using the definition of the inner product in Section II,

$$f^T \hat{\Lambda} S = \langle f, \Lambda S(\cdot) \rangle + \sum_{t \in T_f} f(t)^T (M_0(t) S(t) - M_1(t) S(t\gamma_1) - M_2(t) S(t\gamma_2)) + \sum_{t \in T_f} f(t)^T (M_0(t) S(t) - M_1(t) S(t\gamma_1)) \quad (A.2)$$

and

$$(\hat{\Lambda} f)^T S = \langle \Lambda f, S(\cdot) \rangle + \sum_{t \in T_f} f(t)^T M_0(t) S(t) - (f(t\gamma_1))^T M_{-}(t\gamma_1) + f(t\gamma_2)^T M_{-}(t\gamma_2) S(t) + \sum_{t \in T_f} f(t)^T M_0(t) S(t) - f(t)^T M_1(t)^T S(t). \quad (A.3)$$

Therefore, since we have $f^T \Lambda S = (\Lambda f)^T S$, we obtain, by using (A.2) and (A.3),

$$\begin{aligned} \langle f, \Lambda S(\cdot) \rangle - \langle \Lambda f, S(\cdot) \rangle &= \sum_{t \in T_f} f(t)^T (M_1(t) S(t\gamma_1) + M_2(t) S(t\gamma_2)) \\ &\quad - \sum_{t \in T_f} (f(t\gamma_1))^T M_{-}(t\gamma_1) + f(t\gamma_2)^T M_{-}(t\gamma_2) S(t) \\ &\quad + \sum_{t \in T_f} f(t)^T M_{-}(t) S(t) - f(t)^T M_1(t)^T S(t) \end{aligned} \quad (A.4)$$

as in (9).

Now we show that the Green's function $\Gamma(t, s)$ is symmetric, i.e., $\Gamma(t, s) = \Gamma(s, t)^T$, for $t, s \in T$. It is easy to see that for t or

$$\hat{\Lambda} = \begin{bmatrix} M_0(0) & -M_1(0) & -M_2(0) & 0 & 0 & 0 & 0 \\ -M_{-}(1) & M_0(1) & 0 & -M_1(1) & -M_2(1) & 0 & 0 \\ -M_{-}(2) & M_0(2) & 0 & 0 & 0 & -M_1(2) & -M_2(2) \\ 0 & -M_{-}(3) & 0 & M_0(3) & 0 & 0 & 0 \\ 0 & -M_{-}(4) & 0 & 0 & M_0(4) & 0 & 0 \\ 0 & 0 & -M_{-}(5) & 0 & 0 & M_0(5) & 0 \\ 0 & 0 & -M_{-}(6) & 0 & 0 & 0 & M_0(6) \end{bmatrix}. \quad (A.1)$$

$\in T_i, T_f$, we have $\Gamma(t, s) = 0 = \Gamma(s, t)^T$. Define now the two square matrices Λ and Γ having rows and columns associated with all nodes of $\text{int}(T)$. Each element of the matrix Λ is taken from the same corresponding row and column of Λ , and therefore Λ is a submatrix of Λ , representing only the nodes of $\text{int}(T)$. Γ is on the row associated with node t and the column associated with node s , as element $\Gamma(t, s)$. It can be easily verified, using $\Lambda \Gamma(t, s) = I \delta(t, s)$, $t, s \in \text{int}(T)$ that $\Lambda \Gamma = I$ and $\Lambda^T = \Lambda$ so that $I = (\Lambda \Gamma)^T = \Gamma^T \Lambda^T = \Gamma^T \Lambda$, and therefore $I = \Gamma^T$ and thus $\Gamma(t, s) = \Gamma(s, t)^T$ for $t, s \in \text{int}(T)$ which was left to show.

APPENDIX B

In this appendix, we show that the noise covariance matrices $\mathbb{W}(t)$, $t \in \text{int}(T)$ introduced in Section III are positive definite. Let us assume that not every noise covariance matrix $\mathbb{W}(t)$, $t \in \text{int}(I)$ is positive definite. We scan each scale of the tree above T_f from bottom to top until we find a node s such that $\mathbb{W}(s)$ is not positive definite. Then we stop and we denote the set of nodes scanned so far (including s but excluding T_f) by I_s . The fact that $\mathbb{W}(t) > 0$ for $t \in T \setminus s$ and the construction of $\mathbb{W}(t)$ through the difference equation (19) guarantees that $\mathbb{W}(t)^{-1}$ exists for all $t \in T$. Denote with ϵ the vector containing all $\epsilon(t)$, $t \in T$. Define by Q_s^{-1} the matrix consisting of zeros except for the matrices $\mathbb{W}(t)^{-1}$ on the diagonal. We have

$$E = F[\epsilon, \epsilon^T] = Y_s^T Q_s^{-1} Y_s, \quad (B 1)$$

where Y_s is a lower triangular matrix defined by (16) having all 1's on its diagonal. The matrix F is the covariance matrix of part of the residual process of the reciprocal process, and since we assumed $R > 0$ we must have $E > 0$ as well; otherwise a linear combination of $S(t)$, $t \in \text{int}(T)$ would be identically zero, contradicting the fact that $R > 0$. Therefore $\forall u, u^T I_s u > 0$. Since we assumed that $\mathbb{W}(s)$ is not positive definite $\exists \epsilon_s$ such that $\epsilon_s^T \mathbb{W}(s)^{-1} \epsilon_s \leq 0$. However, since Y_s is a lower triangular matrix with 1's on its diagonal, we can always find u such that

$$Y_s^T u = [0 \quad 0 \quad \epsilon_s^T \quad 0 \quad 0]^T \quad (B 2)$$

resulting in

$$\begin{aligned} u^T I_s u &= [0 \quad 0 \quad \epsilon_s^T \quad 0 \quad 0] \\ &\quad Q_s^{-1} [0 \quad 0 \quad \epsilon_s^T \quad 0 \quad 0]^T \\ &= \epsilon_s^T \mathbb{W}(s)^{-1} \epsilon_s \leq 0 \end{aligned} \quad (B 3)$$

contradicting $E > 0$. Therefore we must have $\mathbb{W}(t) > 0$ for all $t \in \text{int}(I)$ which is what we wanted to show.

REFERENCES

- [1] M. Basseville, A. Benveniste, K. C. Chou, S. A. Golden, R. Nikoukhah, and A. S. Willsky, "Modeling and estimation of multiresolution stochastic processes," *IEEE Trans Inform Theory*, vol. 38, pp. 766-784, Mar. 1992.
- [2] K. C. Chou, A. S. Willsky, and A. Benveniste, "Multiscale recursive estimation, data fusion and regularization," *IEEE Trans Automat Cont*, vol. 39, pp. 464-478, Mar. 1994.
- [3] R. W. Dijkerman, "Multi resolution models of stochastic processes," Ph.D. dissertation, INRS Télécommunications, Université du Québec, Sep. 1994.
- [4] R. W. Dijkerman and R. R. Mazumdar, "Wavelet representations of stochastic processes and multiresolution stochastic models," *IEEE Trans Signal Processing*, vol. 42, pp. 1640-1652, July 1994.

- [5] C. D. Greene and B. C. Levy, "Some new smoother implementations for discrete time Gaussian reciprocal processes," *Int J Control Syst*, no. 5, pp. 1233-1247, 1991.
- [6] B. C. Levy, "Noncausal estimation for discrete Gauss-Markov random fields," in *Realization and Modeling in System Theory, Proc. MTS-89*, vol. 1, M. A. Kaashoek, J. H. van Schuppen, and A. C. M. Ran, eds., Boston: Birkhäuser, 1989, pp. 13-21.
- [7] B. C. Levy, E. Fabre, and A. Benveniste, "Gauss-Markov processes over tree structured lattices and their Markov random field description," Tech. Rep. UC Davis, Jan. 1993.
- [8] B. C. Levy, R. Frezza, and A. J. Krener, "Modeling and estimation of discrete time Gaussian reciprocal Processes," *IEEE Trans Automat Contr*, vol. 35, pp. 1013-1023, Sept. 1990.
- [9] M. R. Luetgen, "Image processing with multiscale stochastic models," Ph.D. dissertation, Mass Inst Technol, Lab Inform Decision Syst, May 1993.

Comments on the Loop Transfer Recovery

Ahmed Rachid

Abstract—This paper deals with observer-based linear control systems. More precisely, it is shown that Tsui's loop transfer recovery procedure can lead to poor results.

I. INTRODUCTION

It is now well understood that the CORC (combined observer/regulator compensator) problem is not a trivial extension of the linear quadratic regulator (LQR) case which has been widely studied since early 1960's. In particular, it has been shown by Doyle [1] that linear quadratic Gaussian (LQG) regulators have no intrinsic robustness properties and can exhibit poor stability margins contrary to the LQR case which ensures the well known $[1/2, \infty]$ gain margin and phase margins of at least 60 degrees at the plant input [2].

An attempt to recover these stability margins for the CORC case is loop transfer recovery (LTR) method which forces the return difference $F_r(s)$ at the break point λ (Fig. 1) to be identical (in the limit) to the return ratio $T_{r,r}(s)$ at the break point λ corresponding to the full state feedback.

Doyle and Stein [3] have proposed an iterative procedure which asymptotically achieves the LTR by adding a fictitious noise at the input plant before designing the observer using Kalman filter formulas. This procedure drives some observer poles toward stable plant zeros and the rest toward infinity.

More recently, an alternative to Doyle and Stein's approach has been given by Tsui [4], [5], [6] and aims to guarantee some stability margins at the breakpoint λ . Unfortunately, the method proposed by Tsui can lead to quite poor results; in the sequel, his technique is discussed and its limitations are highlighted.

II. PROBLEM STATEMENT

Consider the linear continuous MIMO plant

$$\dot{x} = Ax + Bu$$

Manuscript received August 3, 1993.

The author is with Laboratoire des Systemes Automatiques, 7 rue du Moulin Neuf, 80 000 Amiens, France.
IEEE Log Number 9407216.

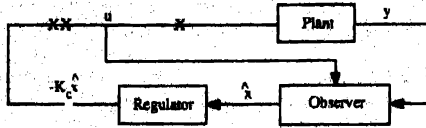


Fig. 1. CORC scheme.

$$y = Cx \quad (1)$$

under the state feedback

$$u = -K_c x. \quad (2)$$

The full observer equations for the considered system are given by

$$\dot{z} = Fz + Ly + TBu \quad (3)$$

under the constraints

$$TA - FT = LC \quad (4)$$

$$K_c = NT. \quad (5)$$

Equivalently, this observer can be described by

$$\dot{\hat{x}} = A\hat{x} + Bu + K_o(y - C\hat{x}).$$

z and \hat{x} are related by $z = T\hat{x}$ from which one can deduce the following relations

$$K_o = T^{-1}L \quad (6)$$

$$A_o = A - K_o C = T^{-1}FT. \quad (7)$$

Fig. 1 summarizes the CORC scheme: At the breakpoint XX , the transfer function can be expressed as

$$\begin{aligned} T_{xx}(s) &= K_c(sI - A_o)^{-1}(B + K_o G(s)) \\ &= K_c(sI - A)^{-1}B \end{aligned} \quad (8)$$

At the breakpoint X , the transfer function is

$$T_x(s) = K_c(sI - A_o + BK_c)^{-1}K_o G(s) \quad (9)$$

where

$$G(s) = C(sI - A)^{-1}B. \quad (10)$$

Doyle and Stein [3] have proved that if

$$K_o[I + C(sI - A)^{-1}K_o^{-1}] = B[C(sI - A)^{-1}B]^{-1} \quad (11)$$

then

$$T_x(s) = T_{xx}(s).$$

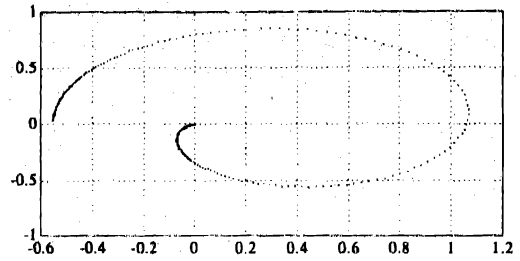
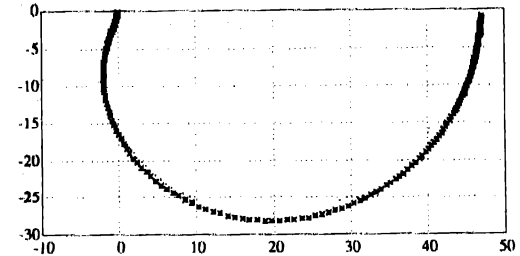
Tsui [4] has proposed the following necessary and sufficient condition allowing LTR when using either Kalman type observer or Lunberger type state or function observers

$$\begin{aligned} T_x(s) &= T_{xx}(s) \text{ if and only if } TB = 0 \text{ or} \\ &N(sI - F)^{-1}TB = 0. \end{aligned}$$

Based on this result, he has proposed a robust observer design procedure which consists in selecting the observer poles so that $\|TB\|$ is minimum. An explicit solution to this optimization problem has also been obtained when the canonical observable form is used for $\{A, B, C\}$.

III. DISCUSSION

First, we notice that Tsui's procedure does not apply to systems with no zeros, i.e., with a constant transfer function numerator; to illustrate this fact, let us consider the following example.

Fig. 2. $T_{xx}(s)$ for the procedure of Tsui.Fig. 3. Full state $T_x(s)$ for Example 2.

Example 1: The considered system is described by

$$G(s) = \frac{3}{(s+1)(s+3)}.$$

The associated observable canonical form (needed in the procedure) is

$$\{A, B, C\} = \left\{ \begin{bmatrix} 0 & -3 \\ 1 & -4 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \end{bmatrix}, [0 \quad 1] \right\}.$$

Let λ_1 and λ_2 be the observer poles to be selected so that TB is small. Then we have two possible cases:

a) $\lambda_1 \neq \lambda_2$:

$$T = \begin{bmatrix} 1 & \lambda_1 \\ 1 & \lambda_2 \end{bmatrix} \rightarrow TB = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$$

b) $\lambda_1 = \lambda_2$:

$$T = \begin{bmatrix} 1 & \lambda_1 \\ 0 & \lambda_2 \end{bmatrix} \rightarrow TB = \begin{bmatrix} 3 \\ 0 \end{bmatrix}.$$

We see that both cases yield to a TB which is independent of the observer poles, so that there is no way to select them using Tsui's procedure. This simple example shows the first gap in the method proposed by Tsui for LTR.

A second major inconvenience in Tsui's technique is that even when it is possible to select the observer poles, LTR is not approximated in many cases. In this context, the following example shows that one can encounter quite unacceptable results.

Example 2: Consider the simple system

$$\{A, B, C\} = \left\{ \begin{bmatrix} 0 & -1 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 0.1 \\ 1 \end{bmatrix}, [0 \quad 1] \right\}.$$

The feedback state controller has been chosen so that $\lambda(A - BK_c) = \{-7, -7\}$ which gives

$$K_c = [-57.78 \quad 17.78].$$

Applying Tsui's procedure, the observer poles λ_1, λ_2 can be chosen as

$$\lambda_1 = -0.1 \quad \text{and} \quad \lambda_2 = -0.11.$$

Fig. 2 shows the corresponding Nyquist of $T_x(s)$ which is to be compared to the Nyquist of the full state feedback transfer function $T_{xx}(s)$ (Fig. 3).

Remark In terms of pole placement Tsui's procedure consists in shifting some of the observer poles at the stable plant zeros and the rest at their neighborhood. This is in contradiction with the Doyle and Stein technique. As a consequence Tsui's technique cannot be applied to nonminimum phase systems contrary to the claims in [6].

REFERENCES

- [1] J. C. Doyle, "Guaranteed margins for IQG regulators," *IEEE Trans Automat Contr*, vol. AC 23, pp. 756-757, 1978.
- [2] M. G. Safonov and M. Athans, "Gain and phase margin for multi-loop IQG regulators," *IEEE Trans Automat Contr*, vol. AC 22, pp. 173-179, 1977.
- [3] J. C. Doyle and G. Stein, "Robustness with observers," *IEEE Trans Automat Contr*, vol. AC 24, pp. 607-611, 1979.
- [4] C. C. Tsui, "On preserving the robustness of an optimal control system with observers," *IEEE Trans Automat Contr*, vol. AC 37, no. 9, pp. 823-828, 1987.
- [5] —, "New approach to robust observer design," *Int J Contr*, vol. 47, no. 3, p. 745, 1988.
- [6] —, "On robust observer compensator design," *Automatica*, vol. 24, no. 5, pp. 687-692, 1988.

Sensitivity Properties of Multirate Feedback Control Systems, Based on Eigenstructure Assignment

R. J. Patton, G. P. Liu, and Y. Patel

Abstract—Some sampled-data systems, e.g., fly-by-wire control schemes, have a necessarily multirate structure, various input and/or outputs sampled at different rates. When considering a multirate system which has parameter uncertainty, it is important to examine ways in which the full freedom of the multivariable design can be utilized to minimize the sensitivity to parameter variations, given the accompanying problems induced by intersample ripple disturbance. This note examines the design capabilities of a class of multirate systems with multiple input and fixed state sampling rates (MIFS), based on eigenstructure assignment. Although the use of eigenstructure assignment for continuous and single rate discrete systems is well understood, the eigenstructure assignment for the design of multirate feedback systems is an open topic of research. Accepting that the problems of intersample ripple are often magnified through multirate control, there are advantages in terms of increased freedom for minimizing sensitivity and optimizing robustness to parameter variations. A special feature of the MIFS class of multirate systems is the ability to introduce extra design freedom in the eigenproblem by a suitable choice of eigenstructure assignment and sample rates. The criteria for the selection of minimum sample rates to produce this extra freedom, and the implication that this has on the eigenstructure assignment problem, are outlined. The improved insensitivity properties are demonstrated using an example comparing the performance of multirate and corresponding single rate designs.

I. INTRODUCTION

For a controllable continuous-time system described by

$$\dot{x}(t) = Ax(t) + Bu(t) \quad (1)$$

Manuscript received March 8, 1991; revised August 23, 1993 and July 7, 1994. This work was supported by the U.K. Engineering and Physical Research Council under Grants GR/H43953.

The authors are with the Department of Electronics, University of York, York YO1 5DD, U.K.

IEEE Log Number 9406998.

where $x \in R^{n \times 1}$ and $u \in R^{m \times 1}$ are the state and control vectors respectively, $A \in R^{n \times n}$, $B \in R^{n \times m}$ and the assignability of a desired self-conjugate set of poles to the equivalent discrete system using multirate feedback control has been well established ([1]-[4] and references therein). For sampled data systems with fast multiple input rates and a slow fixed state rate there are two main problems associated with the application of state feedback control design methods [6]-[7]. The first is the excessive and highly oscillatory nature of state and control effort that is produced. This often makes the implementation of multirate control impractical. The second problem is the increased sensitivity of feedback design to intersample behavior and disturbance effects. The latter is a common drawback of periodic control structures, as highlighted in the work of Francis and Gargios [4]. The problem is particularly prevalent in MIFS control schemes due to fast input control signal updates within a slow fixed state rate. Clearly there is a greater risk of undesirable intersample effects propagating through the system during a slower state period and causing an adverse response before they can be monitored. The two problems are linked: the intersample effects (which cause the system to deviate from its desired behavior) demand a greater control effort to provide the necessary corrective action.

Eigenstructure assignment can be used to alleviate both these problems. The specification of the closed loop system eigenstructure (which comprises the eigenvalues and right eigenvectors) is an important design consideration for the multirate feedback problem. The eigenvectors determine the modal interaction present in the closed loop system transient responses. A multirate feedback design which incorporates a reduction in modal interaction will confine the effects of undesired intersample behavior and disturbances to specific subsystems or individual modes. This minimizes the occurrence of adverse responses due to crossfeed signal which in turn reduces the demanded control effort. Hence a suitable choice of the multirate system eigenvectors will provide an insensitive or well conditioned closed loop performance. Thus the specification of system eigenstructure is seen to address the sensitivity and disturbance issues raised in [4].

The primary motivation for the use of eigenstructure assignment methods to design a multirate system is the ability of the techniques to directly address the problems associated with MIFS feedback control. However a much stronger case for its use is the perfectly decoupled solutions that can be obtained with MIFS sampling. This special characteristic arises from the ability of the MIFS system to generate maximum design freedom for the eigenproblem by an appropriate combination of input sample rates.

II. MIFS MULTIRATE SYSTEM MODEL

The MIFS system of Fig. 1 is classed as an m_1 -input n -state periodic system. The integer m_1 defines the sum of control input updates during a main sample period within which all periodic transitions of the system are described.

To describe the MIFS multirate system where the samplers operate periodically but with different period a complete set of transition equations must be determined for an interval equal to the least common multiple of all sample periods (I in Fig. 1). This interval is referred to as the main interval of sampling. Since the sampling is periodic with I the transition equations will codify the invariant nature of the multirate system.

The MIFS system equations state parameters are represented only at the main interval sample instants, while the control matrix accounts

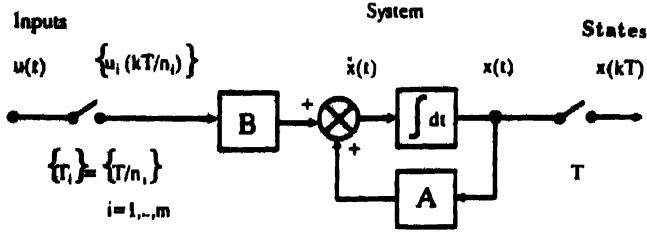


Fig. 1. MIFS sampled multirate system.

for all input changes within the period T . The system sample periods are defined by the following set of relationships:

$$T_b = \frac{T}{n_0}, \quad T_i = \frac{T}{n_i}, \quad m_\mu = \sum_{i=1}^m n_i \quad (2)$$

for $i = 1, 2, \dots, m$, where T_b is the base sample period, n_0 is the least common multiple of the positive integers n_i , and T_i is the i th input sample period.

These sampling interrelationships describe an MIFS sampled system. The time-invariant behavior of the multirate system states of Fig. 1 is described by

$$x((k+1)T) = e^{AT} x(kT) + \sum_{i=1}^m \sum_{j=0}^{n_i-1} \int_0^{T_i} e^{A(T-jT_i-\tau)} B_i d\tau u_i(kT+jT_i) \quad (3)$$

where u_i is the i th control input, B_i denotes the i th column of the original system control matrix B of (1). Therefore, the sampled data multirate model (3) can be rewritten as

$$x(k+1) = \Phi_M(k)x(k) + \Gamma_M(k)u_M(k) \quad (4)$$

where

$$u_M(k) = \begin{bmatrix} u_{M1}(k) \\ u_{M2}(k) \\ \vdots \\ u_{Mm}(k) \end{bmatrix}, \quad u_{Mi}(k) = \begin{bmatrix} u_i(kT) \\ u_i(kT+T_i) \\ \vdots \\ u_i(kT+(n_i-1)T_i) \end{bmatrix} \quad (5)$$

for $i = 1, 2, \dots, m$, $\Phi_M(k) \in \mathbb{R}^{n \times n}$ and $\Gamma_M(k) \in \mathbb{R}^{n \times m_\mu}$ can be obtained from (3), and $x(k) \in \mathbb{R}^{n \times 1}$ and $u_M(k) \in \mathbb{R}^{m_\mu \times 1}$ are the states and the inputs of the MIFS system, respectively.

Fig. 1 illustrates the system structure with which the MIFS system is formed by keeping one fixed state sampling rate. The diagram relates a state-space system.

III. EIGENSTRUCTURE ASSIGNMENT

For the system of (4), the eigenvalue assignment problem is to design a gain matrix F such that the state feedback control

$$u_M(k) = Fx(k), \quad F \in \mathbb{R}^{m_\mu \times n} \quad (6)$$

produces a closed-loop system matrix $(\Phi_M + \Gamma_M F)$ which has a set of desired self-conjugate eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$. Assume (Φ_M, Γ_M) are completely controllable, a range of feedback gain matrices will provide the solution [8]–[11].

Eigenstructure assignment uses the extra degrees of freedom in the underdetermined solution to specify the closed-loop right eigenvectors, $V = [v_1 \ v_2 \ \dots \ v_n]$, corresponding to the desired self-conjugate set $\{\lambda_i\}$. For an insensitive solution, the eigenvectors are chosen to be as mutually orthogonal as possible or to have a specific

modal structure. A measure of the sensitivity of a solution is the conditioning of the closed-loop system right eigenvectors given by [12], [13]

$$\kappa(V) = \|V\|_2 \|V^{-1}\|_2. \quad (7)$$

The eigenstructure assignment problem is to select a set of the desired eigenvalues $\{\lambda_i\}$ and a set of linearly independent right eigenvector matrix V which minimizes $\kappa(V)$ and, furthermore, satisfies the following relation: $(\Phi_M + \Gamma_M F)v_i = v_i \lambda_i$, for $i = 1, \dots, n$, that is,

$$[\Phi_M - \lambda_i I \quad \Gamma_M] \begin{bmatrix} v_i \\ Fv_i \end{bmatrix} = 0. \quad (8)$$

To find a solution, define

$$S_{\lambda_i} = [\Phi_M - \lambda_i I \quad \Gamma_M] \quad (9)$$

and a compatibly dimensioned matrix

$$R_{\lambda_i} = \begin{pmatrix} N_{\lambda_i} \\ M_{\lambda_i} \end{pmatrix} \quad (10)$$

whose columns form the basis for the nullspace of S_{λ_i} . Thus,

$$[\Phi_M - \lambda_i I \quad \Gamma_M] \begin{pmatrix} N_{\lambda_i} \\ M_{\lambda_i} \end{pmatrix} w_i = 0 \quad (11)$$

for any $m_\mu \times 1$ vector, w_i .

It can be deduced from (8) and (11) that

$$v_i = N_{\lambda_i} w_i \quad (12)$$

and if the $\{v_i\}$ are linearly independent set, a real F can be expressed by

$$F = [M_{\lambda_1} w_1 \quad M_{\lambda_2} w_2 \quad \dots \quad M_{\lambda_n} w_n] V^{-1}. \quad (13)$$

For the single rate eigenproblem of (12), an achievable set of eigenvectors is limited by the nullity of the solution space, n_{λ_i} . In general, $m_\mu = m < n$ indicates an underdetermined problem. In this case, a maximum of m elements of the desired eigenvectors can be achieved exactly, and the achievable set of eigenvectors will only approximate the desired set. For the MIFS multirate system, the dimension of R_{λ_i} can be extended beyond that produced by a fixed single rate system, by virtue of the periodic description used for the MIFS sampling scheme.

IV. FEEDBACK DESIGN OF MIFS SAMPLED SYSTEMS

It is well established that the controllability and observability conditions which must be satisfied by the multirate system for the assignability of an arbitrary set of eigenvalues are determined by the choice of input, output, and state sample rates [1]–[7]. These sample parameters have, however, a much more significant role in the eigenstructure assignment procedure; a suitable choice of sample parameters can be used to produce maximum dimension solution space. The extra degrees of freedom generated in this way can be used to ensure a perfectly conditioned solution [14], [15].

The multirate feedback control problem and sensitivity measures relate to the assignment of n eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ designed by a transition matrix which represents the closed-loop system at the main interval sample instants. The set $\{\lambda_i\}$ corresponds to an equivalent set $\{\lambda_{m_i}\}$, representing the faster eigenvalues of the closed-loop system sampled at the based rate T_b . The two sets are related by $\lambda_{m_i} = (\lambda_i)^{n_0}$. The sensitivity measures relate to the n eigenvalues of the MIFS system representation and, therefore, only provide an assessment of system performance at the main sample instants.

The criteria for the selection of the MIFS multirate sample rates have a perfectly conditioned feedback solution are based on the calculation of the extended controllability matrix

$$= [\Gamma_1 \quad \Phi\Gamma_1 \quad \Phi^{2-1}\Gamma_1 \quad \dots \quad \Gamma_m \quad \Phi\Gamma_m \quad \Phi^{2-1}\Gamma_m] \quad (14)$$

where Γ_i is the i th column of the single rate system control matrix evaluated at T , and Φ is the corresponding single rate state transition matrix i.e., $\Phi = \Phi_{AT}$. And define

$$\mu = \sum_{i=1}^m \mu_i \quad (15)$$

The columns of B are selected such that starting from Γ_1 further columns are added until all vectors associated with Γ_1 are linearly independent. The process is continued for all control inputs required for the achievement of (A, B) controllability. When the condition

$$\sum_{i=1}^j \mu_i = n$$

$j < m$ is reached then all subsequent input sample rates have unity input multiplicity i.e. $\mu_i = 1$ for $i = j+1 \dots m$.

For a controllable system if $m_i \geq n$ eigenvalues and eigenvectors can be assigned arbitrarily; if $m_i < n$ the $n - m_i$ entries of each eigenvector cannot be assigned arbitrarily [16]. Thus the indices μ_i , $i = 1 \dots m$ determine the dimension of the maximum solution subspace that the i th input is capable of generating with an MIFS samples system [17]. A choice of input sample rate $\{T_i\} = \left\{\frac{T}{\mu_i}\right\}$ i.e. $n - \mu$ will produce full design freedom for the eigenstructure assignment procedure of Section III if $\mu \geq n$. This choice of input rates is referred to as the *ideal sample set*. (Note also that a rearrangement of columns in the original control matrix may produce a different set $\{\mu_i\}$.)

If $\mu > n$ then $R_{\lambda_i} \in \mathbb{R}^{(n-\mu_i) \times \mu_i}$ for the eigenproblem of (10) thus providing full design freedom for the eigenstructure assignment procedure. This effectively allows the precise assignment of any set of finite magnitude desired eigenvectors of the closed loop multirate system. For $\mu = n$ the design vector u is determined uniquely resulting in a simpler formula for the multirate gain matrix

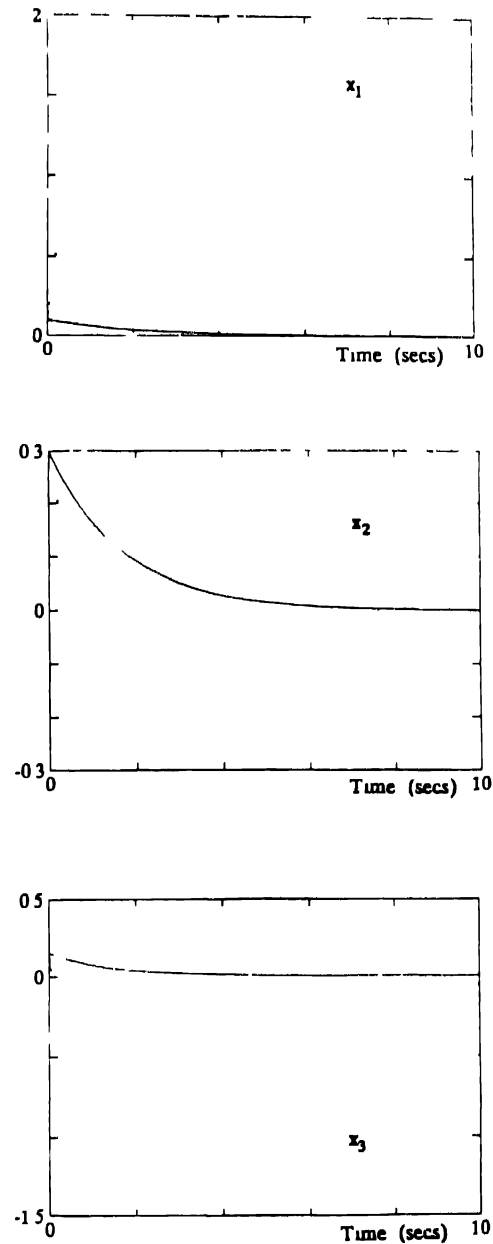
$$K = [V_{\lambda_1}(\lambda_{\lambda_1})^{-1} \dots V_{\lambda_n}(\lambda_{\lambda_n})^{-1}]V^{-1} \quad (16)$$

For $\mu < n$ one method of solving the state feedback eigenproblem is by a least squares minimization of the error between the desired and achieved eigenvectors i.e. the minimization of $\|v_i - \bar{v}_i\|_2$. The least squares projection of the desired eigenvector v_i onto the solution space λ_{λ_i} determines the design vector to be

$$u = (\lambda_{\lambda_i}^T \lambda_{\lambda_i})^{-1} \lambda_{\lambda_i}^T v_i \quad (17)$$

The least squares projection method is termed the 'direct' assignment due to its noniterative assignment of the desired eigenvectors. The direct method is well suited for comparison purposes due to its single pass calculation. This leaves no opportunity for varying levels of attention in the design of the control laws (as for example with optimized assignment methods).

A choice of the 'nonideal' set of sampling rates ($\mu < n$) can also be implemented to design multirate feedback control that provides superior insensitivity performance than that achievable by the corresponding single rate design. This flexibility in the choice of input sample rate design multiplicity allows insensitivity properties of the closed-loop MIFS multirate system to be determined by the designer. An advantage of a 'nonideal' set of sample rates is the improved intersample behavior that can be obtained by effectively relaxing the eigenproblem solvability conditions as demonstrated in Section V.



(a)

Fig. 2 (a) State responses of the closed loop systems formed by the feedback gain matrices K_1 , K_2 (and K_3). — indicates K_1 system, --- indicates K_2 and K_3 systems. (b) Control input responses of the closed loop system with K_2 (multirate). (c) Control input responses of the closed loop system with K_1 (single rate).

For cases where the sample rate selection process gives $\mu > n$, a characteristic of the resulting feedback gain matrix produced is the generation of $\mu - n$ null rows. This is a direct consequence of the overdetermined solution produced by this choice of sample rates. This effectively limits the number of nonzero gain elements of a feedback control matrix of any multirate scheme satisfying the necessary conditions for multirate eigenvalue assignability to $\leq n^2$. The extra design freedom is illustrated in Fig. 2 which shows the additional effective inputs which arise as a consequence of the multisampling.

Eigenstructure assignment techniques are also noted for their ability to generate closed-loop systems which are insensitive to variations or perturbations in the nominal system dynamics [13]. This insensitivity

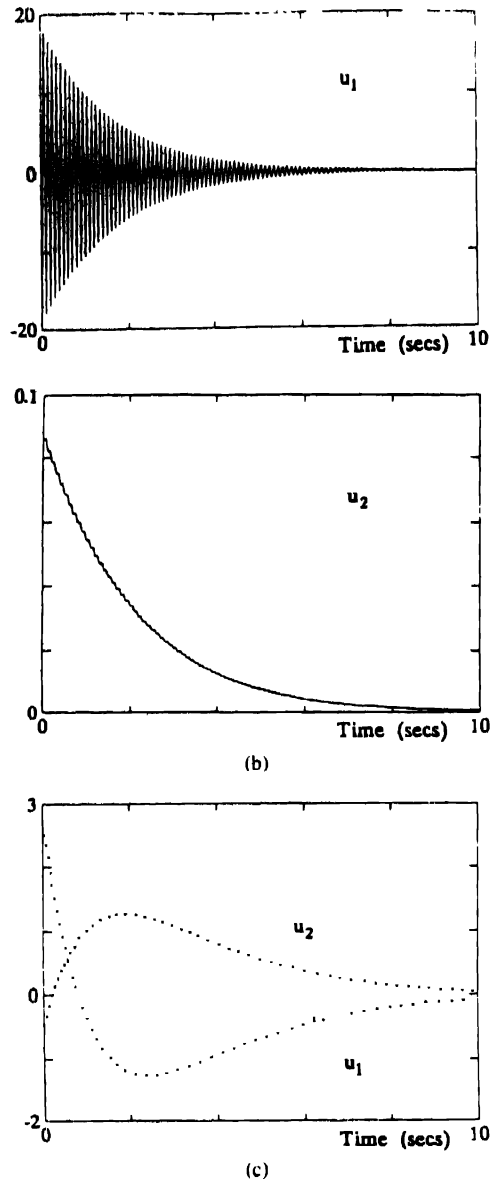


Fig. 2 (b) Control input responses of the closed-loop system with K_2 (multirate). (c) Control input responses of the closed-loop system with K_1 (single rate).

is achieved by decoupling the modal interactions presented in the closed-loop system. The benefits of closed-loop system insensitivity can be observed very clearly in the time domain.

V. EXAMPLE

A simple example is used to demonstrate the effects of applying the direct eigenstructure assignment procedure and the sample rate selection criteria outlined in Sections III and IV. Multirate gain matrices are designed for ideal and nonideal sample rates for the example considered to produce perfectly conditioned solutions to the multirate eigenstructure assignment problem. The performance of the multirate feedback designs is compared to the corresponding single rate design to demonstrate the improved insensitivity of the multirate designs.

Let the matrices of a system be

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}. \quad (18)$$

TABLE I

$\kappa(V)$ AND $\|K\|_2$ OF K_1 , K_2 , AND K_3 CLOSED-LOOP SYSTEMS

Performance Measure	Feedback Gain Matrix		
	K_1	K_2	K_3
$\kappa(V)$	172.14	1	1
$\ K\ _2$	7.8253	67.156	75.622

The desired closed-loop eigenvalues $\{-0.5, -0.6, -0.7\}$ are assigned using a least squares solution such that the corresponding right eigenvectors are maximally orthogonal (ideally $\{v_i\} = I_n$). The main interval of sampling is selected to be $T = 0.1$. This does not render any of the modes uncontrollable, thus ensuring the assignability of the desired eigenvalues. Formulating matrix β , gives two choices of input multiplicities: $\{\mu_1 = 2, \mu_2 = 1\}$ and $\{\mu_1 = 3, \mu_2 = 1\}$. The multirate feedback design produced by these two choices of sample rates are given below together with the corresponding single rate design.

- 1) The input sampling intervals $T_1 = T$, $T_2 = T$ and the state sampling interval is T (single rate); this gives the single rate gain matrix K_1

$$K_1 = \begin{pmatrix} 2.9832 & 5.9881 & 3.7679 \\ -0.2931 & -1.0622 & -1.1507 \end{pmatrix}$$

$$\|K_1\|_2 = 7.8253.$$

- 2) The input sampling intervals $T_1 = T/2$, $T_2 = T$, the state sampling interval is T (multirate); this gives the "nonideal" multirate feedback gain matrix K_2

$$K_2 = \begin{pmatrix} 22.1843 & -40.2434 & -16.5515 \\ 20.1605 & 37.5406 & 17.6343 \\ 0.525 & 0.3829 & 0.5409 \end{pmatrix}$$

$$\|K_2\|_2 = 67.1560.$$

- 3) The input sampling intervals $T_1 = T/3$, $T_2 = T$, the state sampling interval is T (multirate); this gives the ideal multirate feedback gain matrix K_3

$$K_3 = \begin{pmatrix} 25.3670 & 45.8164 & 18.3884 \\ 0.0 & 0.0 & 0.0 \\ -22.3352 & -41.7645 & -20.0087 \\ -0.5243 & -0.3821 & 0.5389 \end{pmatrix}$$

$$\|K_3\|_2 = 75.6217$$

which has one null row since $\mu - n = 1$.

Table I shows the eigenvector conditioning $\kappa(V)$ and $\|K\|_2$ figures for the above designs. The $\|K\|_2$ figure gives an indication of the magnitude of the control effort demanded by each design.

Fig. 2 shows the responses of the K_1 and K_2 closed-loop systems to initial state perturbation of $x(0) = [0.1, 0.3, 0.15]^T$. The main sample responses of the K_2 and K_3 design are identical. Fig. 2 shows that both multirate systems exhibit the perfect modal decoupling that all designs are attempting to achieve.

Both multirate feedback designs have a very high $\|K\|_2$ compared to the equivalent single rate design. Thus, an increased demand on the control inputs is predicted (as verified by the responses of Fig. 2). The only difference in the control input responses of the two multirate systems (in addition to the different update rates) is a slight increase in the magnitude of u_1 demanded by feedback design K_1 . The u_2 control signal for all designs is, however, smooth and has an acceptable magnitude. Thus, the faster u_1 sampling required for the

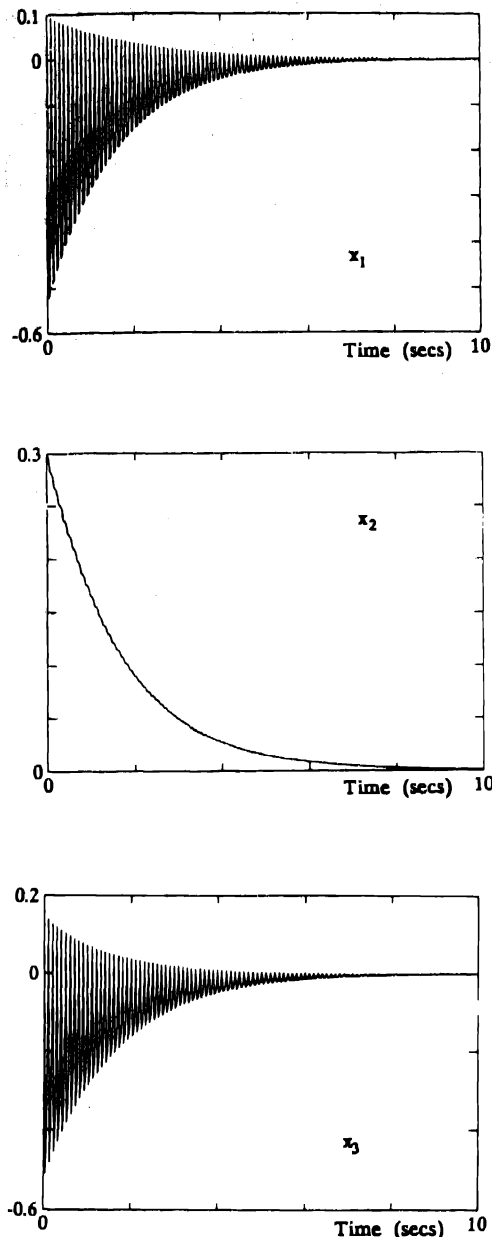


Fig. 3. Intersample state behavior of closed-loop systems formed by the multirate gain matrix K_2 .

K_3 closed-loop system appears to offer no advantage over the K_2 system. Both achieve the required decoupled responses at the main sample instants.

Fig. 2 shows the system responses at a rate $1/T$ (which is the correct state sample rate of the MIFS system). The state responses during the interval T are shown in Fig. 3. These responses show that the improved decoupling of the multirate designs is obtained at the cost of large magnitude, switched intersample state behavior (x_2 of both multirate design does not suffer from any intersample switching).

The most important point to note from the tabulated performance measure and the time responses is the achievement of perfect modal decoupling by multirate feedback gains K_2 and K_3 at the main sample instants. In comparison, the modal decoupling achievable by the single rate design K_1 is relatively poor.

The insensitivity of the multirate and single rate closed-loop systems to perturbations in the nominal open loop systems dynamics

TABLE II
THE CLOSED-LOOP NOMINAL AND PERTURBED EIGENVALUES

Nominal Eigenvalues	Perturbed Eigenvalues	
	K_1	K_2 and K_3
0.9512	1.0153	$0.9530 + j0.0070$
0.9324	0.9136	$0.9530 - j0.0070$
0.9418	0.9504	0.9364

is tested. A comparison of the shift in position of the K_1 (single rate), K_2 and K_3 closed-loop system poles caused by this perturbation will establish the sensitivity of each system.

The sensitivities of the closed-loop systems produced by K_1 , K_2 , and K_3 are tested by perturbing the open-loop system transition matrix A . A proportional perturbation in the system dynamics of the form $A = (1 + \delta)A$ is considered. A perturbation $\delta = -0.1$ (which corresponds to a 10% perturbation) just makes the single rate system unstable. The closed-loop eigenvalues produced by each feedback gain matrix in response to this perturbation in the nominal system dynamics are given in Table II (together with the nominal eigenvalues).

Thus, the multirate closed-loop system maintains stable performance, while the single rate system becomes unstable. The multirate system can, in fact, tolerate an extreme proportional perturbation in the open-loop system dynamics. A value of $\delta = -0.75$ (i.e., a 75% perturbation) will cause the above K_2 , K_3 systems to become unstable.

VI. CONCLUSION

Many systems have a sampled-data structure which is necessarily multirate to fit in with the use of different data buses and digital sensors and actuators. The criteria for the selection of input sample rate multiplicities to achieve a perfectly conditioned closed-loop system have been outlined. This note has demonstrated the application of eigenstructure assignment to multirate systems. The criteria for the selection of input sample rate multiplicities to achieve a perfect conditioned closed-loop system have been outlined. An example has demonstrated the achievement under certain conditions of perfectly decoupled solutions at the main sample instants by an MIFS system, whose input sample rates are chosen using the proposed criterion. For the example system, two input rate multiplicities (determined using the general selection criterion) were examined. The nominal responses of both multirate designs have been compared to the corresponding single rate designs. The multirate solutions are seen to provide much more insensitive feedback control than that produced by the equivalent single rate design. The results demonstrate the applicability of the eigenstructure assignment technique for MIFS feedback control design.

A direct eigenstructure assignment method, where the desired eigenvectors are approximated using a least squares projection, has been used to solve the multirate eigenstructure assignment problem. This serves to demonstrate the principles of multirate feedback design using eigenstructure assignment. The direct approach is, by no means, the ideal method of utilizing the design freedom offered by MIFS system models. A more useful approach is to optimize the solution to the MIFS eigenproblem such that intersample switching is alleviated while maintaining a near-perfectly conditioned modal structure. An MIFS eigenstructure assignment technique of this type is presented in [14]. The optimization procedure [15] makes full use of the extra MIFS system design freedom to balance the different closed-loop system qualities.

ACKNOWLEDGMENT

Thanks are expressed to J. Chen of University of York for helpful comments.

REFERENCES

- [1] T. Hagiwara and M. Araki, "On the necessary condition for discrete time pole assignability by piecewise constant output feedback," *Int J Contr*, vol. 43, pp. 1905-1909, 1986.
- [2] A. B. Chammass and C. T. Leonides, "On the design of linear time invariant systems by periodic state feedback," *Int J Contr*, vol. 27, pp. 885-894, 1978.
- [3] P. P. Kargonekar, K. Polla, and A. Tannenbaum, "Robust control of linear time invariant plants using periodic compensation," *IEEE Trans Automat Contr*, vol. AC 30, pp. 1088-1096, 1985.
- [4] B. A. Francis and T. T. Gorgious, "Stability theory for linear time invariant plants with periodic digital controllers," *IEEE Trans Automat Contr*, vol. AC 33, pp. 820-832, 1988.
- [5] M. Araki, "Recent developments in digital control theory," preprint of 12th IFAC World Congr., vol. 9, pp. 251-260, Sydney, July 1993.
- [6] T. Hagiwara, T. Fujimura, and M. Araki, "Generalized multirate output controllers," *Int J Contr*, vol. 52, pp. 597-612, 1990.
- [7] Y. Patel and R. J. Patton, "A robust approach to multirate controller design using eigenstructure assignment," in *Proc 1990 Amer Contr Conf*, San Diego, CA, 1990, pp. 945-951.
- [8] B. C. Moore, "On the flexibility offered by state feedback in multivariable systems beyond closed loop eigenvalue assignment," *IEEE Trans Automat Contr*, vol. AC 21, pp. 659-692, 1976.
- [9] G. P. Liu and R. J. Patton, "Parametric state feedback design of multivariable control systems using eigenstructure assignment," in *Proc 32nd IEEE Conf Decision Contr*, San Antonio, TX, Dec. 1993, pp. 835-836.
- [10] ———, "Eigenstructure assignment toolbox for use with Matlab," Dep. Electron. Univ. York, UK, 1994.
- [11] R. J. Patton, G. P. Liu, and J. Chen, "Multivariable control using eigenstructure assignment and the method of inequalities," in *Proc 2nd Euro Contr Conf*, Groningen, The Netherlands, 1993, pp. 1143-1148.
- [12] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, London: Oxford University Press, 1965.
- [13] J. Kautsky, N. K. Nichols, and P. Van Dooren, "Robust pole assignment in linear state feedback," *Int J Contr*, vol. 41, pp. 1129-1155, 1985.
- [14] Y. Patel and R. J. Patton, "Design of robust multirate feedback control using eigenstructure assignment," in *Mathematics of the Analysis and Design of Process Control*, P. Borne and S. G. Izatostas, Eds., 1992.
- [15] S. P. Burrows and R. J. Patton, "Design of a low sensitivity minimum norm and structurally constrained control law using eigenstructure assignment," *Op Contr Appl Methods*, vol. 12, pp. 131-140, 1991.
- [16] S. Srinathkumar, "Eigenvalue/eigenvector assignment using output feedback," *IEEE Trans Automat Contr*, vol. AC 23, pp. 79-81, 1978.
- [17] M. Kono, T. Suzuki, and T. Morishita, "Block decoupling of linear ω -periodic discrete time systems," *IEEE Trans Automat Contr*, vol. 35, pp. 1262-1265, 1990.

Optimality Conditions for Truncated Kautz Networks with Two Periodically Repeating Complex Conjugate Poles

Tomás Oliveira e Silva

Abstract—We present optimality conditions for the approximation of an SISO system by a truncated Kautz network with two repeating complex conjugate poles. Both the continuous and the discrete time cases are discussed. We approach the problem in a system approximation framework, and we do not assume that the input signal is white (or an impulse). The results we obtain generalize the results already known for truncated Laguerre networks.

1 INTRODUCTION

The usage of orthonormal series of exponential functions in the representation of signals pioneered by the work of Lee and Wiener in the 1930's [1] is an important topic of research in the areas of automatic control and digital signal processing. For the continuous time (CT) case these orthonormal functions can be obtained quite easily if we express them by their Laplace transforms [1]–[4]

$$G_k(s) = \sqrt{p_k + p_k^*} \frac{\prod_{l=1}^k \frac{1}{s - p_l^*}}{\prod_{l=1}^k \frac{1}{s + p_l}} \quad k \in \mathbb{N} \quad (1)$$

The poles of these functions, which correspond to the exponents of the exponential functions, are arbitrary except that they must have negative real parts, i.e., $\text{Re } p_k < 0$. For the discrete time (DT) case the orthonormal sequences can also be obtained quite easily if we express them by their Z transforms [5]

$$G_k(z) = \sqrt{1 - a_l a_l^*} \frac{\prod_{l=1}^k \frac{1}{1 - a_l^* z^*}}{\prod_{l=1}^k \frac{1}{1 - a_l z}} \quad l \in \mathbb{N} \quad (2)$$

The poles of these functions are arbitrary except that they must be inside the unit circle, $|a_l| < 1$. Of all the possible selections of the poles of these functions, the special case where all the poles are equal (and real), corresponding to the so called Laguerre functions (CT: $p_k = p$) and sequences (DT: $a_l = a$) has attracted the interest of several researchers [6]–[11]. Because in practice we are forced to truncate the Laguerre series expansion of a given or estimated function (sequence) after a finite number of terms, an important problem related to this approach is the choice of a good value for p (or a) [12]–[15]. A directly related problem is the choice of the optimal value of p (or a) for a truncated Laguerre series with a specified number of terms [16]–[18].

Unfortunately, the approximation of a function whose transform has one or more pairs of complex conjugate dominant poles by a truncated Laguerre series requires a large number of terms [9]. A simple way to ameliorate this problem is to use the orthonormal functions given by (1) or (2). To avoid too many degrees of freedom one usually assumes that the poles of the expansion are just two complex conjugate periodically repeating poles, i.e., $p_{k+1} = p$, $p_k = p^*$, $l \in \mathbb{N}$ (and a similar definition for the a_k 's). In this

Manuscript received October 19, 1993; revised June 23, 1994.

The author is with the signal processing group of INESC/Aveiro, Universidade de Aveiro, 3800 Aveiro, Portugal.
IEEE Log Number 9406999.

¹We will denote the real part of a variable by the subscript *re* and its imaginary part by *im*. For Laplace/Z transforms a subscript *re* will denote the transform of the real part of the function/sequence, not the real part of the transform, likewise for *im*. The square root of -1 will be denoted by j .

we can apply an unitary transformation to (1) to obtain [2], [3]

$$G_{2k-1}(s) = \sqrt{p+p^*}(s-|p|) \frac{[(s-p)(s-p^*)]^{k-1}}{[(s+p)(s+p^*)]^k} \quad (3)$$

and

$$G_{2k}(s) = \sqrt{p+p^*}(s+|p|) \frac{[(s-p)(s-p^*)]^{k-1}}{[(s+p)(s+p^*)]^k} \quad (4)$$

For the DT case, (2) can also be replaced by [5]

$$G_{2k-1}(z) = |1+a| \frac{\sqrt{1-aa^*}}{\sqrt{2}} (z^{-1}-1) \frac{[(z^{-1}-a)(z^{-1}-a^*)]^{k-1}}{[(1-az^{-1})(1-a^*z^{-1})]^k} \quad (5)$$

and

$$G_{2k}(z) = |1-a| \frac{\sqrt{1-aa^*}}{\sqrt{2}} (z^{-1}+1) \frac{[(z^{-1}-a)(z^{-1}-a^*)]^{k-1}}{[(1-az^{-1})(1-a^*z^{-1})]^k} \quad (6)$$

We will call these functions (and their inverse transforms) generically by the name of Kautz functions with two periodically repeating complex conjugate poles, or Kautz(cc) for short.

Like the Laguerre functions, the Kautz(cc) functions can approximate arbitrarily well, in the integrated squared error (ISE) sense, any impulse response with finite energy. This property of Laguerre functions was used in the past to model plants with unknown or unmodeled dynamics [6], [7], [19], [20]. Unfortunately, the Laguerre functions are not well suited to approximate functions with strong oscillatory behavior—a task that can be done more efficiently with Kautz(cc) functions. For this reason, some researchers have recently turned their attention to this kind of functions [21], [22].

The problem of deducing the optimality conditions for p (or a) in the approximation, in the ISE sense, of the impulse response of an SISO plant by a truncated Laguerre series was solved in [16] and [17]. Recently, these results have been extended to the more general case where the plant to be identified is excited by a signal with a nonconstant power spectrum and where the plant is approximated by a truncated Laguerre network [18]. We will extend here the results of [18] to the case of a truncated Kautz(cc) network. To this end, we will first deduce the optimality conditions for a truncated Laguerre network with one complex pole. We then proceed to show that the real part of the output of this network is the output of a Kautz(cc) network. Finally, using some results of the Laguerre case, we deduce the optimality conditions for Kautz(cc) networks. We conclude the note with a simple numerical example.

A. CT versus DT Signals and Deterministic versus Stationary Stochastic Signals

It is interesting to note that (5) and (6) are quite similar to (3) and (4). This suggests that the results for DT are similar to those for CT, which in fact is the case. We will therefore discuss in detail only the CT case. The DT case will be dealt with in a couple of remarks.

Another point worth mentioning is the similarity of the results for deterministic and for stationary stochastic signals. In the rest of this note, we will need to evaluate inner products of the form (CT, $s = j\omega$)

$$(FX, GY) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(j\omega) G^*(j\omega) X(j\omega) Y^*(j\omega) d\omega$$

where F , G , X , and Y are the Laplace transforms of deterministic signals with finite energy. When $x(t)$ and $y(t)$ are stationary stochastic signals with finite power, we can also evaluate an inner product of this form by replacing in the above expression $X(j\omega)Y^*(j\omega)$ by the generalized Fourier transform of the cross-correlation between $x(t)$ and $y(t)$. Therefore, by a suitable modification of the inner product definition, we can cover the cases of deterministic and stochastic signals simultaneously.

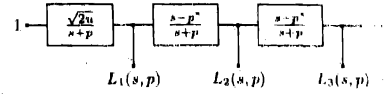


Fig. 1. Network for generating the continuous-time Laguerre functions with one complex pole. (The variable u is the real part of p .) A similar network exists to generate the discrete-time Laguerre sequences.

II. OPTIMALITY CONDITIONS FOR TRUNCATED LAGUERRE NETWORKS WITH ONE COMPLEX POLE

We will basically follow the approach of [18], adapted to our present needs and to the CT case. For convenience, we will assume that $p = u + jv$. The CT Laguerre functions are obtained easily from (1), with the result

$$L_k(s, p) = \sqrt{2u} \frac{(s-p^*)^{k-1}}{(s+p)^k}, \quad k \in \mathbb{N}.$$

These functions can be easily generated by the block diagram of Fig. 1. By applying to this network a general input signal, and by forming a linear combination of its various outputs to approximate a given desired signal, we obtain a Laguerre network.

Let $X(s)$ be the Laplace transform of the input signal of the Laguerre network, and let $D(s)$ be that of the desired signal. Both signals are assumed to have finite energy and to be real. Let n be the number of sections of the Laguerre network, and let $Y_n(s, p)$ be the Laplace transform of the network's output. We then have

$$Y_n(s, p) = \sum_{k=1}^n w_{n,k} L_k(s, p) X(s).$$

Note the dependence of the weights $w_{n,k}$ on the number of sections. This is a consequence of the fact that the L_k 's are not orthonormal when the weighting function, the power spectrum of the input signal, is not the constant function 1. With the notation $A_k(s, p) = L_k(s, p) X(s)$, the error signal is given by

$$E_n(s, p) = \sum_{k=1}^n w_{n,k} A_k(s, p) - D(s)$$

and, by denoting the usual inner product of complex signals by (\cdot, \cdot) , we obtain the ISE of the Laguerre network with n sections as $\xi_n = (E_n, E_n)$. (Due to Parseval's theorem, we have chosen to represent the ISE as an inner product in the transform domain.)

The optimal values for the weights $w_{n,k}$ can be easily found by applying the orthogonality property of Hilbert spaces, the result being the following set of so-called normal equations:

$$(E_n, A_k) = 0, \quad k = 1, \dots, n. \quad (7)$$

Furthermore, we have for the partial derivative of ξ_n with respect to a generic real parameter (a prime denotes here differentiation with respect to that parameter):

$$\xi'_n = \sum_{k=1}^n [w_{n,k} (A'_k, E_n) + w_{n,k}^* (E_n, A'_k)]. \quad (8)$$

To actually compute the partial derivatives with respect to u and v , we will need the following remarkable formulas:

$$\frac{\partial L_k(s, p)}{\partial u} = \frac{(1-k)L_{k-1}(s, p)}{2u} + k \frac{L_{k+1}(s, p)}{2u} \quad (9)$$

$$\frac{\partial L_k(s, p)}{\partial v} = j \frac{(k-1)L_{k-1}(s, p) + (1-2k)L_k(s, p)}{2u} + jk \frac{L_{k+1}(s, p)}{2u}. \quad (10)$$

Because we have $A'_k = L'_k \lambda$, we conclude that²

$$\frac{\partial A_k}{\partial u} = k \frac{A_{k+1}}{2u} + \text{terms in } A_k \text{ and } A_{k-1} \quad (11)$$

and that

$$\frac{\partial A_k}{\partial v} = jk \frac{A_{k+1}}{2u} + \text{terms in } A_k \text{ and } A_{k-1} \quad (12)$$

Applying (11) and (12) to (8) and simplifying the result with (7), we obtain for the stationary points of ξ_n with respect to p (i.e., with respect to u and v) the following system of equations

$$\begin{cases} u_{n+1} (A_{n+1}, E_n) + u_{n+1}^* (E_n, A_{n+1}) = 0 \\ u_{n+1} (A_{n+1}, E_n) - u_{n+1}^* (E_n, A_{n+1}) = 0 \end{cases}$$

It is a simple matter to verify that the only solutions of this system are

$$u_{n+1} = 0 \quad \text{or} \quad (E_n, A_{n+1}) = 0$$

We will now show that the second condition is equivalent to $u_{n+1} = 0$. To this end, we apply the Gram-Schmidt orthonormalization procedure to the set $\{A_k\}$, obtaining the orthonormal set $\{B_k\}$. This can always be done because there are no linear dependencies between the A_k 's.³ We can express the orthonormalization procedure by

$$B_n = T_n A_n$$

where T_n is a lower triangular non-singular matrix. A_n is a column matrix defined by

$$A_n = [A_1 \quad A_2 \quad \dots \quad A_n]^T$$

and B_n is defined in a similar way. To simplify the notation in the rest of this section, we will denote the element of the last row and column of the matrix T_n by r_n . Note that $r_n \neq 0$ and that for $\lambda(s) = 1$ we have $r_n = 1$ for all n . It is also interesting to note that the element of the last row and column of the matrix T_n^{-1} is equal to r_n^{-1} .

Due to the orthonormality of the B_k 's, we have

$$F_{n+1} = E_n + r_{n+1} B_{n+1}$$

where $r_{n+1} = (D, B_{n+1})$ is the $n+1$ th Fourier coefficient associated with the expansion of $D(s)$ with respect to the orthonormal set $\{B_k\}$. Therefore, we have

$$(E_n, A_{n+1}) = (E_{n+1}, A_{n+1}) - r_{n+1} (B_{n+1}, A_{n+1})$$

One of the normal equations for a network with $n+1$ sections gives us immediately $(E_{n+1}, A_{n+1}) = 0$. We also have $(B_{n+1}, A_{n+1}) = r_{n+1}^{-1} \neq 0$. From these facts we conclude that the condition $(E_n, A_{n+1}) = 0$ is equivalent to $r_{n+1} = 0$. We finish the proof by noting that $u_{n+1} = r_{n+1} = 0$.

From the above results, we conclude that the stationary points of the truncated Laguerre network satisfy

$$u_{n+1} u_{n+1}^* = 0 \quad (13)$$

as well as the normal equations (7) for orders n and $n+1$. Equation (13) is exactly the optimality condition obtained in [16] and [17] under much more restrictive assumptions. Because the normal equations define completely the weights $u_{n,k}$ (and the weights $u_{n+1,k}$), the solutions of (13) are the values of p for which $u_{n,n} = 0$ or $u_{n+1,n+1} = 0$. One of these solutions will correspond to the global minimum of ξ_n (other solutions may correspond to local minima or maxima, or even to saddle points).

²The term in A_k in (11) is missing for the CT case but is present for the DT case [cf. (14)].

³In fact, by factoring out the common nonzero term $\lambda(s)$ an operation that can always be done because $\lambda(s)$ is an analytic function in the right-half plane, the resultant set is orthonormal. (Obviously, we are not considering the uninteresting case $\lambda(s) = 0$.)

Remark 1 For the DT case, we have for the Z -transforms of the Laguerre sequences

$$L_k(z, a) = \sqrt{1 - a^2 - a^2} \frac{(z^{-1} - a^*)}{(1 - az^{-1})^k}, \quad k \in \mathbb{N},$$

where $a = u + jv$, and the remarkable formulas

$$\frac{\partial L_k(z, a)}{\partial u} = \frac{(1-k)L_{k-1}(z, a) + j(1-2k)L_k(z, a)}{1 - a^2 - a^2} + k \frac{L_{k+1}(z, a)}{1 - a^2 - a^2} \quad (14)$$

$$\frac{\partial L_k(z, a)}{\partial v} = j \frac{(k-1)L_{k-1}(z, a) + (2k-1)L_k(z, a)}{1 - a^2 - a^2} + jk \frac{L_{k+1}(z, a)}{1 - a^2 - a^2} \quad (15)$$

Note that the last term of (14) is almost equal to that of (9) — the same happening between (15) and (10). For this reason, (11) and (12) are also valid for the DT case (with $2u$ replaced by $1 - a^2 - a^2$). Therefore, the proof for the CT case can be adapted quite easily to the DT case.

III. OPTIMALITY CONDITIONS FOR TRUNCATED SECOND-ORDER KAUTZ NETWORKS WITH TWO COMPLEX CONJUGATE POLES

In this section we will show that the real part of the output of the Laguerre network of the previous section is the output of a Kautz network of second order with two complex conjugate poles and with the same number of sections as the Laguerre network.⁴ We then proceed to compute the optimality conditions for truncated Kautz networks. To simplify the exposition we will discuss here only the CT case. The DT case is similar.

The Laplace transform of the real part of the output of a Laguerre network with a complex pole is given by⁵

$$\begin{aligned} Y_{re}(s, p) &= \frac{\lambda(s)}{2} \sum_{k=1}^n [u_{n,k} L_k(s, p) + u_{n,k}^* L_k(s, p^*)] \\ &= \sqrt{\frac{u}{2}} \lambda(s) \sum_{k=1}^n F_{n,k}(s, p) \end{aligned}$$

where

$$F_{n,k}(s, p) = u_{n,k} \frac{(s - p^*)^{k-1}}{(s + p)^k} + u_{n,k}^* \frac{(s - p)^{k-1}}{(s + p^*)^k}$$

It is clear that this last expression can be put in the form

$$F_{n,k}(s, p) = \sum_{m=1}^k (\phi_{n,k,m} s + \psi_{n,k,m}) \frac{[(s + - + p^*)(s - p)]^{n-1}}{[(s + p)(s + p^*)]^n}$$

where all coefficients involved in this expression are real.⁶ It is then obvious that $Y_{re}(s, p)/\lambda(s)$ can be expressed in a unique way as a linear combination of equations (3) and (4), i.e., $Y_{re}(s, p)$ is the output of a second-order Kautz network. More specifically, we have

$$Y_{re}(s, p) = \sum_{k=1}^n [a_{n,k} G_{-k-1}(s, p) + b_{n,k} G_{2k}(s, p)] \lambda(s)$$

where the $a_{n,k}$'s and the $b_{n,k}$'s can be obtained from the $u_{n,k}$'s (or the $\phi_{n,k,m}$'s and $\psi_{n,k,m}$'s) by an invertible transformation. Because

⁴Note that for p (or a) real (i.e., $v = 0$), the Kautz network collapses into a Laguerre network (with twice the number of sections). We will not discuss this special case here.

⁵Note that the input signal is real.

⁶It can be shown that the conditions $\phi_{n,k,m} = 0$ and $\psi_{n,k,m} = 0$ are equivalent to the condition $u_{n,k} = 0$ when $v \neq 0$.

and $b_{r,n}$ are related to $\phi_{r,n}$ and $\psi_{r,n}$ by a invertible linear transformation when $\epsilon \neq 0$, we conclude that

$$\begin{cases} a_{r,n} = 0 \\ b_{r,n} = 0 \end{cases} \Leftrightarrow \begin{cases} \phi_{r,n} = 0 \\ \psi_{r,n} = 0 \end{cases} \Rightarrow u_{r,n} = 0 \quad (16)$$

We now turn our attention to the problem of deducing the optimality conditions for truncated second order Kautz networks. The starting point is the equation

$$J = Y_{rr} - D = \sum_{k=1}^n [u_{r,k} A_{r,k} - u_{rr,k} A_{rr,k}] - D$$

where we have explicitly separated the real from the imaginary parts of all complex variables and functions. The ISE to be minimized is now $\xi_r = (F_r, F_{rr})$. The subscript r in ξ is just to distinguish the Kautz case from the Laguerre case. Note that we have

$$A_{r,k}(s, p) = \frac{A_k(s, p) + A_k(s, p^*)}{2}$$

and

$$A_{rr,k}(s, p) = \frac{A_k(s, p) - A_k(s, p^*)}{2j}$$

From these equations and from (9) and (10) it is possible to show that

$$\frac{\partial A_{r,k}}{\partial u} = \frac{A_{r,k+1}}{2u} + \text{other terms} \quad (17)$$

$$\frac{\partial A_{rr,k}}{\partial u} = k \frac{A_{rr,k+1}}{2u} + \text{other terms} \quad (18)$$

$$\frac{\partial A_{r,k}}{\partial t} = -I \frac{A_{r,k+1}}{2u} + \text{other terms} \quad (19)$$

$$\frac{\partial A_{rr,k}}{\partial t} = k \frac{A_{rr,k+1}}{2u} + \text{other terms} \quad (20)$$

where the "other terms" are terms in $A_{r,j}, A_{rr,k}, A_{r,k+1}$ and $A_{rr,k+1}$ (some may be missing).

In the present case the weights $u_{r,j}$ have to satisfy the following set of normal equations

$$\begin{cases} (F_r, A_{r,k}) = 0 \\ (F_{rr}, A_{rr,k}) = 0 \end{cases} \quad k = 1, \dots, n \quad (21)$$

Note that these normal equations are *different* from those in (7). In fact in the Kautz case we have $2n$ equations with $2n$ *real* unknowns while in the Laguerre case we have n equations in n *complex* unknowns. Therefore the set of weights for the optimal Laguerre case when transformed to the Kautz case usually do not produce an optimal solution.

The partial derivatives of ξ_r with respect to a real parameter u are now given by

$$\begin{aligned} \xi_r &= 2(F_r, F_{rr}) \\ &= 2 \sum_{k=1}^n [u_{r,k} (F_r, A_{r,k}) - u_{rr,k} (F_{rr}, A_{rr,k})] \end{aligned}$$

Using (17) to (20) in this expression and simplifying the result with (1) we obtain the following stationary conditions for ξ_r ,

$$\begin{cases} u_{r,k} (F_r, A_{r,k+1}) - u_{rr,k} (F_{rr}, A_{rr,k+1}) = 0 \\ u_{r,k} (F_{rr}, A_{rr,k+1}) + u_{rr,k} (F_r, A_{r,k+1}) = 0 \end{cases}$$

The only solutions of this system of equations are

$$u_{r,k} = 0 \quad \text{or} \quad \begin{cases} (F_r, A_{r,k+1}) = 0 \\ (F_{rr}, A_{rr,k+1}) = 0 \end{cases}$$

As in the Laguerre case we will now show that the second condition is equivalent to $u_{r,k+1} = 0$. We start by orthonormalizing

the set $\{A_{r,k}, A_{rr,k}\}$ obtaining the set $\{B_{r,k}, B_{rr,k}\}$. This can be expressed by

$$B_{r,k} = T_{r,k}^{-1} A_{r,k}$$

where again we have used the subscript r to denote the Kautz case.⁷ Again $T_{r,k}$ is a lower triangular nonsingular matrix which in this case can be partitioned in 2×2 blocks. We will denote by R the 2×2 block whose elements are the elements of the last two rows and columns of $T_{r,n}$. Note that R is itself a lower triangular nonsingular matrix. The 2×2 block similarly positioned in the matrix $T_{r,n}^{-1}$ is equal to R^{-1} .

Using the orthonormal set we have⁸

$$F_{r,n+1} = F_r + \epsilon_{r,n+1} B_{r,n+1} + \epsilon_{rr,n+1} B_{rr,n+1}$$

Using this expression and after some manipulations similar to the Laguerre case we obtain⁹

$$\begin{bmatrix} (F_r, A_{r,n+1}) \\ (F_{rr}, A_{rr,n+1}) \end{bmatrix} = -R^{-1} \begin{bmatrix} \epsilon_{r,n+1} \\ \epsilon_{rr,n+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Because $|R| \neq 0$ the only solution of this system is $\epsilon_{r,n+1} = 0$ which implies $u_{r,n+1} = 0$. As in the Laguerre case we conclude that the stationary points for the Kautz case satisfy

$$u_{r,n+1} = u_{rr,n+1} = 0$$

as well as the normal equations (21) for orders n and $n+1$. According to (16) these conditions can be translated to weights of the Kautz network as

$$\begin{cases} a_{r,n+1} = 0 \\ b_{r,n+1} = 0 \end{cases} \quad \text{or} \quad \begin{cases} a_{r,n+1} + b_{r,n+1} = 0 \\ b_{r,n+1} - a_{r,n+1} = 0 \end{cases} \quad (22)$$

as long as $\epsilon \neq 0$.

Remark 2. The corresponding proof for the DT case is in every respect similar to that of the CT case. In particular we have (for $\epsilon \neq 0$)

$$Y_r(z, a) = \sum_{k=1}^n [a_{r,k} G_{r,k}(z, a) + b_{r,k} G_{rr,k}(z, a)] V(z)$$

where $a_{r,k} = b_{r,k} = 0$ if and only if $u_{r,k} = 0$. Equations (17)–(20) remain valid in the DT case with $2u$ replaced by $1 - u - \epsilon$.

IV. EXAMPLE

To illustrate the results of the previous section we will approximate the rational system

$$H(z) = \sum_{k=1}^n 1 - \frac{1}{k} z^{-k} \quad (23)$$

by a truncated Kautz network using as input the signal

$$V(z) = \frac{1}{1 + 0.8z^{-1}} \quad (24)$$

The values of all parameters of this system are presented in Table I.

The normalized ISL surface of the approximation of the output of (23) by the output of a Kautz network with four sections ($n = 4$) is

⁷The vector $A_{r,n}$ is now defined as

$$A_{r,n} = [A_{r,n}, A_{rr,n}, A_{r,n}, A_{rr,n}]^T$$

⁸For convenience we will represent the two real Fourier coefficients as the real and imaginary parts of a complex number.

⁹Note that

$$R^{-1} = \begin{bmatrix} (B_{r,n+1}, A_{r,n+1}) & (B_{rr,n+1}, A_{r,n+1}) \\ (B_{r,n+1}, A_{rr,n+1}) & (B_{rr,n+1}, A_{rr,n+1}) \end{bmatrix}$$

TABLE 1
PARAMETERS OF THE SYSTEM (23). THESE PARTICULAR VALUES
CORRESPOND TO A FOURTH-ORDER ELLIPTIC BANDPASS FILTER

k	z_k	τ_k
1	0	0.1074
2	$0.5819 + 0.7487j$	$-0.003028 - 0.05312j$
3	$0.5819 - 0.7487j$	$-0.003028 + 0.05312j$
4	$0.7640 + 0.5814j$	$-0.006167 + 0.04034j$
5	$0.7640 - 0.5814j$	$-0.006167 - 0.04034j$

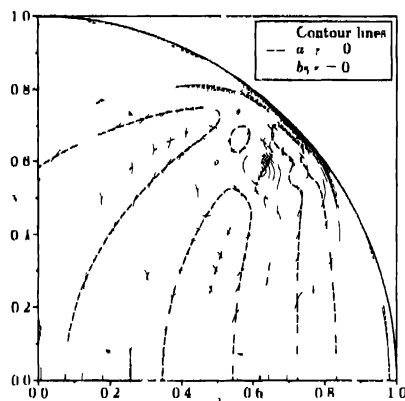


Fig. 2 Contour plot of the normalized ISE of the approximation of the output of (23) by the output of a Kautz network with four sections as a function of the pole position (only first quadrant shown). The input of both systems was (24). Superimposed in this figure are the curves $a_{1,4} = 0$ and $b_{1,4} = 0$. For each value of p , the weights are computed from the normal equations.

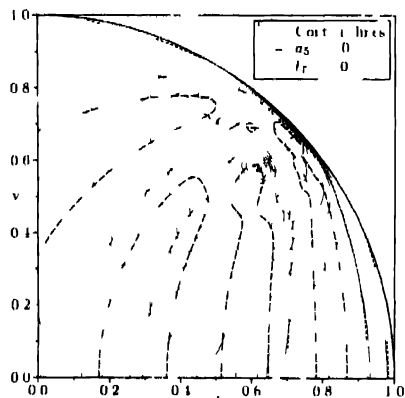


Fig. 3 Same as Fig. 2 but with the curves $a_{5,5} = 0$ and $b_{5,5} = 0$ instead of $a_{1,4} = 0$ and $b_{1,4} = 0$. Note that the global minimum is located at a point ($p \approx 0.619 + j0.580$) where the two curves cross.

displayed in Figs. 2 and 3 using contour lines separated by 1 dB¹⁰. The input of both systems was (24). We have superimposed the curves $a_{1,4} = 0$ and $b_{1,4} = 0$ in Fig. 2, and the curves $a_{5,5} = 0$ and $b_{5,5} = 0$ in Fig. 3. Note that, in this example, the global minimum is located at a point where $u_{5,5} = 0$ ($p \approx 0.619 + j0.580$). However, our conditions do not exclude the possibility that $u_{n,n} = 0$ (and not $u_{n+1,n+1} = 0$) may hold at a global minimum. This is known to be possible in Laguerre networks.

¹⁰For each value of p , we have computed the weights from (21). For this reason, in these figures the normalized ISE is only a function of p .

V CONCLUSIONS

We have deduced the optimality conditions for truncated Kautz networks with two complex conjugate periodically repeating poles. It is shown that these conditions are a simple generalization of those of the Laguerre case. The system approximation case discussed here is more general than the function approximation case usually considered in the literature. Surprisingly, the optimality conditions are the same in the two cases.

Another proof of these results, which is not based on a Laguerre network with a complex pole and is therefore easier to generalize, is presented in [23] (only for the special case of an impulsive input).

REFERENCES

- [1] Y. W. Lee, *Statistical Theory of Communication*, New York: Wiley, 1960.
- [2] W. H. Kautz, 'Transient synthesis in the time domain', *IRE Trans. Circuit Theory*, vol. 1, pp. 29-39, 1954.
- [3] D. C. Ross, 'Orthonormal exponentials', *IEEE Trans. Commun. Electron.*, pp. 173-176, Mar. 1964.
- [4] J. M. Mendel, 'A unified approach to the synthesis of orthonormal exponential functions useful in systems analysis', *IEEE Trans. Syst. Sci. Cybern.*, vol. 2, pp. 54-62, Aug. 1966.
- [5] P. W. Broome, 'Discrete orthonormal sequences', *J. Assoc. Comput. Mach.*, vol. 12, no. 2, pp. 151-168, Apr. 1965.
- [6] C. C. Zervos and G. A. Dumont, 'Deterministic adaptive control based on Laguerre series representation', *Int. J. Control*, vol. 48, no. 6, pp. 2333-2359, 1988.
- [7] G. A. Dumont, C. C. Zervos, and G. I. Pagueau, 'Laguerre based adaptive control of pH in an industrial bleach plant extraction stage', *Automatica*, vol. 26, no. 4, pp. 781-787, 1990.
- [8] P. M. Mäkilä, 'Approximation of stable systems by Laguerre filters', *Automatica*, vol. 26, no. 2, pp. 333-345, 1990.
- [9] B. Wahlberg, 'System identification using Laguerre models', *IEEE Trans. Automat. Contr.*, vol. 36, pp. 551-562, May 1991.
- [10] S. Gunnarsson and B. Wahlberg, 'Some asymptotic results in recursive identification using Laguerre models', *Int. J. Adaptive Control. Signal Processing*, vol. 5, pp. 313-333, 1991.
- [11] A. C. den Brinker, 'Adaptive modified Laguerre filters', *Signal Processing*, vol. 31, pp. 69-79, 1993.
- [12] E. E. Ward, 'The calculation of transients in dynamical systems', *Proc. Cambridge Philosophical Soc.*, vol. 50, pp. 49-59, 1954.
- [13] T. W. Parks, 'Choice of time scale in Laguerre approximations using signal measurements', *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 511-514, Oct. 1971.
- [14] M. Schetzen, 'Asymptotic optimum Laguerre series', *IEEE Trans. Circuit Theory*, vol. CT-18, pp. 493-500, Sept. 1971.
- [15] Y. Fu and G. Dumont, 'An optimum time scale for discrete Laguerre network', *IEEE Trans. Automat. Contr.*, vol. 38, pp. 934-938, June 1993.
- [16] G. J. Clowes, 'Choice of the time scaling factor for linear system approximations using orthonormal Laguerre functions', *IEEE Trans. Automat. Contr.*, vol. AC-10, pp. 487-489, Oct. 1965.
- [17] M. A. Masnadi Shirazi and N. Ahmed, 'Optimal Laguerre networks for a class of discrete time systems', *IEEE Trans. Signal Processing*, vol. 39, pp. 2104-2108, Sept. 1991.
- [18] T. Oliveira e Silva, 'Optimal conditions for truncated Laguerre networks', *IEEE Trans. Signal Processing*, vol. 39, Sept. 1994.
- [19] C. Zervos, P. R. Bélanger, and G. A. Dumont, 'On PID controller tuning using orthonormal series identification', *Automatica*, vol. 24, no. 2, pp. 165-175, 1988.
- [20] O. Agamennoni, E. Paolini, and A. Dasages, 'On robust stability analysis of a control system using Laguerre series', *Automatica*, vol. 28, no. 4, pp. 815-818, 1992.
- [21] P. Lindskog and B. Wahlberg, 'Applications of Kautz models in system identification', in *Preprints 12th IFAC World Cong.*, July 1993, pp. 309-312.
- [22] Y. Fu and G. A. Dumont, 'On determination of Laguerre filter pole through step or impulse response data', in *Preprints 12th IFAC World Cong.*, July 1993, pp. 303-307.
- [23] A. C. den Brinker, F. P. A. Benders, and T. Oliveira e Silva, 'Optimality conditions for truncated Kautz series,' *IEEE Trans. Circuits Syst.* submitted for publication, 1994.

Stable Adaptive Control of a Class of First-Order Nonlinearly Parameterized Plants

Jovan D. Bošković

Abstract—Stable adaptive control algorithms are designed for a class of first-order nonlinearly parameterized plants containing nonlinearities similar to those arising in fermentation process models. When the plant nonlinearity is conveniently parameterized and when the controller parameters are adjusted adaptively, the output error is shown to converge to zero. Stability of the overall system is proved using the Lyapunov function with a suitably chosen cubic term.

1 INTRODUCTION

The past several years have witnessed substantial progress in the field of adaptive feedback linearizing control. While most of the existing results are derived in the context of linear parameterizations, in [6] the control problem is also solved for a class of nonlinearly parameterized plants, both in nonadaptive and adaptive cases. In this note we suggest a method for solving the adaptive control problem for a class of nonlinearly parameterized models similar to those arising in fermentation processes. Unlike the adaptive method from [6] which is based on high gain adaptation, in this note we suggest the method based on certainty equivalence principle and a novel form of Lyapunov functions.

The class of plants considered in this paper contains a nonlinearity which is of the form of a ratio of polynomials in x with three unknown parameters, the two of which do not enter linearly. Such a model captures all the important features of a large class of models of fermentation processes. It is shown that, for a constant desired response of the plant, the appropriate plant parameterization, in conjunction with the corresponding adaptive algorithms and a suitably chosen Lyapunov function, enables us to prove, without neglecting any terms in the derivative of the Lyapunov function, that the output error of the plant converges to zero, i.e., that $\lim_{t \rightarrow \infty} x(t) = 0$. The results presented in this note are based on the results from [3] and are aimed to be an important step towards designing stable adaptive algorithms for a class of uncertain nonlinear plants arising in continuous flow and fed batch fermentation processes.

The paper is organized as follows. In Section II the motivation is given and the tracking problem is stated for a special class of nonlinear plants. In Section III two plant parameterizations are presented and the corresponding error models are derived. Section IV contains the stability analysis for both cases of plant parameterizations, while the conclusions are given in Section V.

II MOTIVATION AND PROBLEM STATEMENT

To illustrate the difficulties encountered when attempting to control a nonlinear fermentation process, we will consider a class of fermentation process models in which the growth of microorganisms and substrate consumption are described by the Monod's growth model. In this case, using the notation adopted in biochemical engineering, the process model can be described by

$$\dot{\lambda} = \lambda \left[\frac{\mu_m S}{K_s + S} - D \right] \quad (1)$$

$$\dot{S} = -\frac{1}{Y} \frac{\mu_m S \lambda}{K_s + S} + (S_f - S)D \quad (2)$$

where λ and S denote respectively microorganisms and substrate concentrations and D denotes the dilution rate, $D = 1/\tau$, where τ denotes the substrate feed and V denotes the volume of the fermentation broth ($V = \text{const}$ for continuous flow processes and $V = F\tau$ for fed-batch processes). In the above equations, μ_m denotes the maximum value of the specific growth rate, K_s denotes the Monod's coefficient, Y denotes the yield coefficient, and S_f denotes the concentration of the substrate in the feed.

In the above equations the functional form of the growth model may be known, while the coefficients μ_m , K_s and Y may vary in an unknown fashion within known bounds. If we wanted to regulate $S(t)$ around some desired value S^* , we would use a control law of the form

$$D(t) = \frac{1}{S_f - S(t)} \left\{ \frac{1}{Y} \frac{\mu_m(t) S(t) \lambda(t)}{K_s(t) + S(t)} - \lambda[S(t) - S^*] \right\}$$

where $\lambda > 0$. In the case we wanted to regulate $\lambda(t)$ around λ^* , the control law would take the form

$$D(t) = \frac{1}{\lambda(t)} \left\{ \frac{\mu_m(t) S(t) \lambda(t)}{K_s(t) + S(t)} + \lambda[\lambda(t) - \lambda^*] \right\}$$

The question naturally arises as to how should the parameter estimates be adjusted in each case to provide that the output assumes values close to the desired one over an interval of interest. Due to the fact that some of the unknown parameters enter the static equation nonlinearly, the way in which the estimates should be adjusted so that the control objective is achieved is not immediately obvious. Thus a question arises whether the process nonlinearities can be more conveniently parameterized in a fashion which would simplify the analysis and the design of adaptive algorithms. In this context we will address two problems:

- (1) how to conveniently parameterize the nonlinearities from (1)
- (2) which would simplify the design of adaptive laws, and
- (2) how to prove the stability of the overall system.

To solve these problems we adopt a prototype plant whose nonlinearity captures main features of the nonlinearities from (1). The prototype plant is chosen in the form

$$\dot{x} = \frac{\rho f_1(x)}{k + k_1 x} + q(t)u - x(t) \quad (3)$$

where $x(t)$ represents a physical quantity in total coordinates, such that it cannot assume values less than zero. The input and the output of the plant are respectively denoted by $u(t)$ and $y(t)$ and $u \in \mathcal{S}$, where \mathcal{S} is a suitably chosen set. Also we assume that $t \in \mathbb{R}^+$ and that $|q(t)| > 0 \forall t \in (0, +\infty)$. It is further assumed that the value of ρ can be either +1 or -1. We now introduce the parameter vector $p = [f_1, k, k_1]^T$, and the set \mathcal{S}_p such that $p \in \mathcal{S}_p = \{p \mid p = [f_1, k, k_1]^T, 0 < (f_1)_{\min} \leq f_1 \leq (f_1)_{\max}, +\infty < k_1 \leq k < +\infty, j = 1, 2, 3\}$. It is also assumed that only $(k_1)_{\min}$ and $(k_1)_{\max}$ ($j = 1, 2, 3$) are known as well as that $t \in \mathcal{S} = \{t \mid 0 \leq t \leq t_f, t_f \in (0, +\infty)\}$.

Comment. The state and parameters of the plant (3) can assume only nonnegative values, which substantially restricts the class of plants under consideration. On the other hand, nonlinear models of fermentation processes, as well as a large number of nonlinear models arising in process industry, are given in total coordinates. In such a case the total values of the process states (e.g., pressures, levels, temperatures, flows, concentrations and other variables) assume strictly nonnegative values. Such models are also characterized by nonnegative values of parameters. These facts may justify the analysis of the restricted class of plants (3).

Manuscript received October 26, 1993; revised July 20, 1994.

The author is with the Center for Systems Science, Department of Electrical Engineering, Yale University, New Haven, CT 06520 USA.
IEEE Log Number 9407236.

We next consider the control objective for the plant (3). We will assume that the desired response of the plant is defined by

$$\dot{x}^*(t) = -\lambda_d x^*(t) + b_m r, x^*(0) = r \quad (4)$$

where $\lambda = b_m \in (0, +\infty)$, $0 \leq r \leq r_{\max} < +\infty$, and r is constant.

We further introduce the output error $e(\cdot) = x(\cdot) - r$. Since $e(0) = e_0 = x_0 - r$, we have that $e_0 \in \mathcal{S}_{e_0} = \{e_0 : -r_{\max} \leq e_0 \leq r_1 + r_{\max}\}$. Now we are ready to state the control objective for the plant (3).

Objective Design a control law for the plant (3) such that, for $p \in \mathcal{S}_p$ and for $e_0 \in \mathcal{S}_{e_0}$, all signals in the system are bounded, and $\lim_{t \rightarrow \infty} e(t) = 0$.

To solve this problem, we will first consider different possible parameterizations of the plant nonlinearity.

III. PLANT PARAMETRIZATIONS AND ERROR MODELS

A Left Parameterization

When $\rho = 1$, we use the left parameterization of the plant (3) [3]. Such a parameterization is of the form

$$\dot{x} = \frac{\alpha_1 x^2}{\alpha_2 + x} + g(x)u \quad (5)$$

where $\alpha_1 = k_1/k_3$ and $\alpha_2 = k_2/k_3$. We further introduce $\alpha = [\alpha_1 \ \alpha_2]^T$, and $\alpha \in \mathcal{S}_\alpha = \{\alpha : \alpha = [\alpha_1 \ \alpha_2]^T, 0 < (\alpha_j)_{\min} \leq \alpha_j \leq (\alpha_j)_{\max}, j = 1, 2\}$, where $(\alpha_j)_{\min} = (k_j)_{\min}/(k_3)_{\max}$, $(j = 1, 2)$, and $(\alpha_j)_{\max} = (k_j)_{\max}/(k_3)_{\min}$, $(j = 1, 2)$. In this case the control law is chosen in the form

$$u(t) = \frac{1}{g(x(t))} \left[-\frac{\theta_1(t)x^2(t)}{\theta_2(t) + x(t)} - \lambda x(t) + b_m r \right] \quad (6)$$

where θ_1 and θ_2 denote the adjustable controller parameters. To avoid division by zero in the above expression (and to assure the stability of the overall system, as will be shown in Section IV), the adaptive algorithms should be designed to ensure that

$$\theta(t) \in \mathcal{S}_\theta = \{\theta : \theta = [\theta_1 \ \theta_2]^T, 0 < (\alpha_j)_{\min} - \epsilon_j \leq \theta_j \leq (\alpha_j)_{\max} + \epsilon_j, j = 1, 2, \forall t \in \mathbb{R}^+\} \quad (7)$$

where $\epsilon_j = \min\{(\alpha_j)_{\min}, 0.5[(\alpha_j)_{\max} - (\alpha_j)_{\min}]\}$, $j = 1, 2$ [1], [3]. After substituting expression (6) into (3) and subtracting (4), we obtain

$$\begin{aligned} \dot{e}(t) = & -\lambda e(t) - \frac{\phi_1(t)x^2(t)}{\theta_2(t) + x(t)} \\ & + \frac{\alpha_1 \phi_2(t)x^2(t)}{[\alpha_2 + r(t)][\theta_2(t) + x(t)]} \end{aligned} \quad (8)$$

where $\phi_j(\cdot) = \theta_j(\cdot) - \alpha_j$, $(j = 1, 2)$, denote the controller parameter errors.

In this case the adaptive algorithms are suggested in the form

$$\begin{aligned} \dot{\phi}_j(t) = \dot{\theta}_j(t) = & (-1)^{j+1} \frac{\gamma_j e(t)x^2(t)}{[\theta_2(t) + x(t)]^j} \\ & \cdot \left\{ 1 - \text{sign}[e(t)x(t)] \cdot \frac{(\alpha_j)_{\max} - 2\theta_j(t) + (\alpha_j)_{\min}}{(\alpha_j)_{\max} + 2\epsilon_j - (\alpha_j)_{\min}} \right. \\ & \cdot \left. \frac{1}{k_{\theta_j}[(\alpha_j)_{\max} - \theta_j(t) - \epsilon_j][\theta_j(t) + (\alpha_j)_{\min} - \epsilon_j] + 1} \right\} \\ j = 1, 2 \end{aligned} \quad (9)$$

where $\gamma_j > 0$, $(j = 1, 2)$ denote the adaptive gains, and where $k_{\theta_j} \gg 1$, $(j = 1, 2)$. The system (8), (9) represents the error model for the case of the left parameterization.

The adaptive algorithms of the form (9) represent "smooth" adaptive algorithms with projection [2, 1]. Such algorithms assure that (7) holds. Also, it can be shown [2, 1] that (for $r = \text{const.}$) the following holds

$$\begin{aligned} \phi_j \dot{\phi}_j \leq & (-1)^{j+1} [\theta_j - \alpha_j] \frac{\gamma_j e x^2}{[\theta_2 + x]^j}, \\ \forall (\theta, \alpha, x) \in & (\mathcal{S}_\theta \times \mathcal{S}_\alpha \times \mathbb{R}^+), j = 1, 2. \end{aligned} \quad (10)$$

B Right Parameterization

In the case when $\rho = -1$, we will use the right parameterization of the plant (3) [3], which takes the form

$$\dot{x} = -\frac{\alpha_1 x^2}{1 + \alpha_2 x} + g(t, r)u \quad (11)$$

where $\alpha_1 = k_1/k_2$ and $\alpha_2 = k_3/k_2$. In this case $\alpha = [\alpha_1 \ \alpha_2]^T$, and $\alpha \in \mathcal{S}_\alpha = \{\alpha : \alpha = [\alpha_1 \ \alpha_2]^T, 0 < (\alpha_j)_{\min} \leq \alpha_j \leq (\alpha_j)_{\max}, j = 1, 2\}$, where $(\alpha_j)_{\min} = (k_{2j-1})_{\min}/(k_2)_{\max}$, $(j = 1, 2)$, and $(\alpha_j)_{\max} = (k_{2j-1})_{\max}/(k_2)_{\min}$, $(j = 1, 2)$. Also, in this case the control law is chosen as

$$u(t) = \frac{1}{g(r(t))} \left[\frac{\theta_1(t)x^2(t)}{1 + \theta_2(t)x(t)} - \lambda x(t) + b_m r \right]. \quad (12)$$

As in the previous case, we want to assure that (7) holds, for the quantities as defined above.

After substituting expression (12) into (3) and subtracting (4) we obtain

$$\begin{aligned} \dot{e}(t) = & -\lambda e(t) + \frac{\phi_1(t)x^2(t)}{1 + \theta_2(t)x(t)} \\ & - \frac{\alpha_1 \phi_2(t)x^2(t)}{[1 + \alpha_2 x(t)][1 + \theta_2(t)x(t)]} \end{aligned} \quad (13)$$

where, as in the previous case, $\phi_j(\cdot) = \theta_j(\cdot) - \alpha_j$, $(j = 1, 2)$, denote the controller parameter errors.

In this case the adaptive algorithms are suggested in the form

$$\begin{aligned} \dot{\phi}_j(t) = \dot{\theta}_j(t) = & (-1)^j \frac{\gamma_j e(t)x^2(t)}{[1 + \theta_2(t)x(t)]^j} \\ & \cdot \left\{ 1 - \text{sign}[e(t)x(t)] \cdot \frac{(\alpha_j)_{\max} - 2\theta_j(t) + (\alpha_j)_{\min}}{(\alpha_j)_{\max} + 2\epsilon_j - (\alpha_j)_{\min}} \right. \\ & \cdot \left. \frac{1}{k_{\theta_j}[(\alpha_j)_{\max} - \theta_j(t) - \epsilon_j][\theta_j(t) + (\alpha_j)_{\min} - \epsilon_j] + 1} \right\}, \\ j = 1, 2 \end{aligned} \quad (14)$$

where $k_{\theta_j} \gg 1$, $(j = 1, 2)$. Equations (13), (14) constitute the error model in the case of the right parameterization.

As in the previous case it can be also be shown that the following holds

$$\begin{aligned} \phi_j \dot{\phi}_j \leq & (-1)^j [\theta_j - \alpha_j] \frac{\gamma_j e x^{j+1}}{[1 + \theta_2 x]^j}, \\ \forall (\theta, \alpha, x) \in & (\mathcal{S}_\theta \times \mathcal{S}_\alpha \times \mathbb{R}^+), j = 1, 2. \end{aligned} \quad (15)$$

IV. STABILITY ANALYSIS

From the linear relationship $\phi_j = \theta_j - \alpha_j$, $(j = 1, 2)$, and from the definitions of the sets \mathcal{S}_α and \mathcal{S}_θ , it follows that

$$\begin{aligned} \phi(t) \in \bar{\mathcal{S}}_\phi = & \left\{ \phi : \phi = [\phi_1 \ \phi_2]^T, |\phi_j| \leq (\alpha_j)_{\max} \right. \\ & \left. - (\alpha_j)_{\min} + \epsilon_j, j = 1, 2, \forall t \in \mathbb{R}^+ \right\}. \end{aligned} \quad (16)$$

We also define sets \mathcal{S}_e and \mathcal{S}_ϕ which respectively represent open connected neighborhoods of points $e = 0$ and $\phi = 0$, with $\mathcal{S}_\phi \supset \bar{\mathcal{S}}_\phi$.

As it is well known [5], on the set $\mathcal{S} \times \mathcal{S}_\omega$ there always exists a function $V(\epsilon, \omega)$ satisfying $V \in C^{(1)}(\mathcal{S}_\epsilon \times \mathcal{S}_\omega)$, and the conditions

$$0 < b_{11}\epsilon^2 + b_{12}\|\omega\|^2 \leq V(\epsilon, \omega) \leq b_{21}\epsilon^2 + b_{22}\|\omega\|^2 \quad \forall (\epsilon, \omega) \in \mathcal{S}_\epsilon \times \mathcal{S}_\omega \quad (17)$$

where $b_{ij} \in (0, +\infty)$, $(i = 1, 2, j = 1, 2)$. The conditions (17) imply that $V(\cdot)$ is positive definite and decrescent on $\mathcal{S} \times \mathcal{S}$, as well as that it is a radially unbounded function.

Now we are ready to state the tracking conditions for the plant (3).

Theorem If the following conditions are simultaneously satisfied i) the conditions (16) and (17) hold, and ii) along the motions of the corresponding error model the following holds

$$\dot{V}(\epsilon, \omega) \leq -\lambda \epsilon^2 \leq 0 \quad \forall (\epsilon, \omega) \in (\mathcal{S} \times \mathcal{S}) \quad \forall t \geq 0 \quad (18)$$

then $\lim_{t \rightarrow \infty} \epsilon(t) = 0$.

Proof The proof follows directly from [7].

The next step is to test the stability conditions in the two cases of plant parameterizations. Since in both cases of plant parameterizations it can be readily shown that a quadratic tentative Lyapunov function is not well suited for the stability analysis of the plant (3) in this paper we will present special tentative Lyapunov functions with cubic terms originally suggested in [3].

A Left Parameterization

In this case the tentative Lyapunov function is of the form

$$V(\epsilon, \omega) = \frac{1}{2} \left(\epsilon^2 + \frac{\omega_1^2}{1} + \alpha_1 \omega^2 \right) + \frac{1}{3} \left(\frac{\alpha_1}{1 + \alpha_1} \right) \omega^3 \quad (19)$$

The first question that arises is whether $V(\epsilon, \omega)$ as defined above satisfies conditions (17). Expression (19) can be rewritten as

$$V(\epsilon, \omega) = \frac{1}{2} \left(\epsilon^2 + \frac{\omega_1^2}{1} + \frac{\alpha_1(\alpha_1 + 3\epsilon + 2\theta)}{3(\alpha_1 + 1)} \omega^3 \right) \quad (20)$$

and since the adaptive algorithms (9) assure that (7) holds, it can be readily shown that (17) holds as well. Also, since (7) holds, this implies that (16) holds as well, so that the condition i) of the Theorem is satisfied. Further, since

$$\dot{V}(\epsilon, \omega) = \epsilon\dot{\epsilon} + \frac{\omega_1\dot{\omega}_1}{1} + \alpha_1 \left(\frac{\theta_2 + \epsilon}{\alpha_1 + 1} \right) \omega\dot{\omega}$$

and with (10) in mind, the first derivative of the function $V(\epsilon, \omega)$ along the motions of (8)–(9) yields

$$\begin{aligned} \dot{V}(\epsilon, \omega) &\leq -\lambda \epsilon^2 - \frac{\epsilon\omega_1\dot{\omega}_1}{\theta + 1} + \frac{\alpha_1\epsilon\omega\dot{\omega}}{(\alpha_1 + 1)(\theta + 1)} \\ &\quad + \frac{\epsilon\omega_1\dot{\omega}_1}{\theta + 1} - \frac{\alpha_1\epsilon\omega_2\dot{\omega}}{(\theta + 1)} - \frac{\theta + 1}{\alpha_2 + 1} \\ &\leq -\lambda \epsilon^2 - \frac{\alpha_1\epsilon\omega_2\dot{\omega}}{(\alpha_1 + 1)(\theta + 1)^2(\alpha_1 + 1)} \\ &\leq -\lambda \epsilon^2 \leq 0 \quad \forall (\epsilon, \omega) \in (\mathcal{S} \times \mathcal{S}) \quad \forall t \geq 0 \end{aligned}$$

which implies that condition ii) of the theorem is satisfied. It can be concluded that when $\rho = 1$ when the control law (6) is used and when its parameters are adjusted according to (9) the output error converges to zero.

B Right Parameterization

In this case the tentative Lyapunov function is chosen in the form

$$V(\epsilon, \omega) = \frac{1}{2} \left(\epsilon^2 + \frac{\omega_1^2}{1} + \alpha_1 \frac{\omega_2^2}{\gamma} \right) + \frac{1}{3\gamma^2} \left(\frac{\alpha_1\gamma}{1 + \alpha_1\gamma} \right) \omega^3 \quad (21)$$

The above expression can be rewritten as

$$V(\epsilon, \omega) = \frac{1}{2} \left(\epsilon^2 + \frac{\omega_1^2}{\gamma} + \frac{\alpha_1(3 + \alpha_1\gamma + 2\theta\gamma)}{3\gamma(1 + \alpha_1\gamma)} \omega^3 \right) \quad (22)$$

and since (7) holds, it follows that (17) holds as well. Also, (7) implies that (16) holds, so that condition i) of the theorem is satisfied. We also have that

$$\dot{V}(\epsilon, \omega) = \epsilon\dot{\epsilon} + \frac{\omega_1\dot{\omega}_1}{\gamma} + \frac{\alpha_1}{\gamma} \left(\frac{1 + \theta\gamma}{1 + \alpha_1\gamma} \right) \omega\dot{\omega}$$

We now evaluate the first derivative of $V(\epsilon, \omega)$ along the motions of the system (13)–(14). Keeping in mind (15) we obtain

$$\begin{aligned} \dot{V}(\epsilon, \omega) &\leq -\lambda \epsilon^2 + \frac{\epsilon\omega_1\dot{\omega}_1}{1 + \theta\gamma} - \frac{\alpha_1\epsilon\omega\dot{\omega}}{(1 + \alpha_1\gamma)(1 + \theta\gamma)} \\ &\quad - \frac{\epsilon\omega_1\dot{\omega}_1}{1 + \theta\gamma} + \frac{\alpha_1\epsilon\omega\dot{\omega}}{(1 + \theta\gamma)} - \frac{1 + \theta\gamma}{1 + \alpha_1\gamma} \\ &\leq -\lambda \epsilon^2 - \frac{\alpha_1\epsilon\omega\dot{\omega}}{(1 + \alpha_1\gamma)(1 + \theta\gamma)^2(1 + \alpha_1\gamma)} \\ &\leq -\lambda \epsilon^2 \leq 0 \quad \forall (\epsilon, \omega) \in (\mathcal{S} \times \mathcal{S}) \quad \forall t \geq 0 \end{aligned}$$

It follows that for $\rho = -1$ when the control law (12) is used and when its parameters are adjusted according to (14) the output error converges to zero.

Comment In both cases of plant parameterizations it can be readily shown that the closed loop system is robust to parameter time variations and bounded external disturbances. The proof of robustness is based on the fact that the adaptive algorithms assure that the controller parameter errors are bounded for all time.

Comment One of the questions that arise is whether it is possible to use left parameterization in the case when $\rho = -1$. In such a case the plant is of the form

$$\dot{x} = -\frac{\alpha_1 x}{\alpha_1 + 1} + q(t)u$$

and the control input is chosen as

$$u(t) = \frac{1}{q(t - \tau(t))} \left[\frac{\theta_1(t)\dot{x}(t)}{\theta(t) + \tau(t)} - \lambda x(t) + b - \dot{x} \right]$$

which leads to the following error model

$$\dot{\epsilon}(t) = -\lambda \epsilon(t) + \frac{\omega_1(t)\dot{\omega}_1(t)}{\theta(t) + \tau(t)} - \frac{\alpha_1\omega_2(t)\dot{\omega}_2(t)}{[\alpha_1 + \tau(t)][\theta(t) + \tau(t)]}$$

In this case the adaptive algorithms are of the form

$$\begin{aligned} \omega_j(t) &= \theta_j(t) = (-1)^j \frac{\dot{x}(t)\dot{x}^*(t)}{[\theta(t) + \tau(t)]^j} \\ &\quad \left\{ 1 - \text{sign}[\epsilon(t)] \frac{(\alpha_1)_j \gamma - 2\theta_j(t) + (\alpha_1)_j \gamma}{(\alpha_1)_j \gamma + 2\epsilon_j - (\alpha_1)_j} \right. \\ &\quad \left. \frac{1}{k_{\theta_j}[(\alpha_1)_j \gamma - \theta_j(t) - \epsilon_j][\theta_j(t) + (\alpha_1)_j \gamma + 1]} \right\} \\ &\quad j = 1, 2 \end{aligned}$$

for which the following holds

$$\begin{aligned} \omega_j\dot{\omega}_j &\leq (-1)^j [\theta_j - \alpha_1] \frac{\dot{x}^2}{[\theta_2 + \tau]^j} \\ \forall (\theta, \alpha_1) &\in (\mathcal{S}_\theta \times \mathcal{S}_\alpha \times \mathbb{R}^+) \quad j = 1, 2 \end{aligned}$$

If we choose the tentative Lyapunov function of the form (19), its derivative being

$$\dot{V}(\epsilon, \omega) = \epsilon\dot{\epsilon} + \frac{\omega_1\dot{\omega}_1}{\gamma} + \frac{\alpha_1}{\gamma} \left(\frac{\theta_2 + \epsilon}{\alpha_1 + 1} \right) \omega\dot{\omega}$$

we obtain

$$\begin{aligned} \dot{V}(r, \phi) &\leq -\lambda e^2 + \frac{e\phi_1 x^2}{\theta_2 + x} - \frac{\alpha_1 e\phi_2 x^2}{(\alpha_2 + x)(\theta_2 + x)} \\ &\quad - \frac{e\phi_1 x^2}{\theta_2 + x} + \frac{\alpha_1 e\phi_2 x^2}{(\theta_2 + x)^2} \cdot \frac{\theta_2 + r}{\alpha_2 + r} \\ &\leq -\lambda e^2 + \frac{\alpha_1 e^2 \phi_2^2 x^2}{(\alpha_2 + x)(\theta_2 + x)^2(\alpha_2 + r)}. \end{aligned}$$

It is seen that the proof of stability goes through if

$$\lambda > \frac{4\phi_{\max}^2(\alpha_1)_{\max}}{27(\alpha_2 + r)}$$

where $\phi_{\max} = (\alpha_2)_{\max} + e_2 - (\alpha_2)_{\min}$, which may be a restrictive requirement. On the other hand, when the right parameterization is applied, the system is stable for any $\lambda > 0$.

V. CONCLUSION

In this paper the stability conditions are proved for a class of first-order plants, which contain nonlinearities similar to those arising in fermentation process growth models. Such nonlinearities are nonnegative functions of the state of the system as well as of the possibly unknown process parameters.

It was shown that different signs in front of the plant nonlinearity lead to different plant parameterizations. Such parameterizations are crucial in the choice of the control laws and adaptive algorithms for controller parameter adjustment. These parameterizations also serve as a guideline for the design of appropriate Lyapunov functions used to prove the stability conditions. The Lyapunov functions $V(\cdot)$ used in the stability analysis contain suitably chosen cubic terms which assure positive definiteness of $V(\cdot)$ as well as that the stability conditions are satisfied. Steps involved in the choice of the plant parameterization, the control laws, adaptive algorithms and the Lyapunov function, form a basis of a method for designing robust adaptive control algorithms for a class of fermentation processes. Further development of such a method is currently in progress [4].

REFERENCES

- [1] J. D. Bošković, "Adaptive control of a class of nonlinear time-varying plants with application to fermentation processes," (in Serbian), Ph.D. dissertation, Univ. Belgrade, Yugoslavia, Jan. 1992.
- [2] —, "Modified adaptive algorithms with projection," in *Proc Sixth Yale Workshop on Adaptive and Learning Systems*, Yale Univ. New Haven, CT, Aug. 1990.
- [3] —, "Parameter adaptive tracking control of a class of first order nonlinear plants," in *Proc 1993 IEEE/SMC Conference*, Le Touquet, France, 1993.
- [4] —, "Stable adaptive control of a class of bioreactor processes," Center for Systems Science, Yale Univ., Tech. Rep. 1994.
- [5] W. Hahn, *Theory and Application of Lyapunov's Direct Method* (International Series in Applied Mathematics) Englewood Cliffs, NJ: Prentice-Hall, 1963.
- [6] R. Marino and P. Tomei, "Global adaptive output-feedback control of nonlinear systems, part ii: nonlinear parameterization," *IEEE Trans Automat. Contr.*, vol. 38, pp. 33–48, Jan. 1993.
- [7] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems* Englewood Cliffs, NJ: Prentice-Hall, 1988.

Multiproduct Production/Inventory Control Under Random Demands

Jin Qiu and Richard Loulou

Abstract—In this note we study the optimal production/inventory control policy for a single machine multiproduct production system. The machine produces to fill the end-product inventory stock and the demand is satisfied from the inventory when available; unsatisfied demand is backlogged until the product becomes available as the result of production. For each product, the demand follows a Poisson process and the unit processing time is known. When the machine switches production from one product to another, it incurs a set-up time and a set-up cost. The relevant costs include the set-up cost, a cost per unit time while the machine is running, and linear costs for inventory and backlogging.

This problem is modeled as a semi-Markov decision process using the criterion of minimizing expected total cost with discounting over an infinite horizon. Procedures for computing near-optimal policies and their error bounds are developed. The error bound given by our procedure is shown to be much tighter than the one given by the "Norm-based" approach. Computational test results are presented to show the structure of the near-optimal policy and how its accuracy is affected by the system characteristics such as capacity utilization and set-up time.

I. INTRODUCTION

There has been a growing interest in design of multiproduct production/inventory control policies for capacitated manufacturing systems. A good control policy leads to low inventory and backlogging costs. Policies studied frequently in the literature of operations management include base stock policies (Karmarkar [8], Graves [7]), lot-sizing policy (Karmarkar *et al.* [9], Zipkin [17]), and linear control rules (Graves [5], Denardo, and Tang [2]). These policies are heuristic in nature since they are restricted to a predetermined policy structure (e.g., constant production lot size in lot-sizing policies).

In this note we explore the optimal production/inventory control policy for a single machine multiproduct production system. The machine produces to fill the end-product inventory stock, and the demand is satisfied from the inventory when available; unsatisfied demand is backlogged until the product becomes available as the result of production. There is limited stock space for each product. When the machine switches production from one product to another, it incurs a set-up time and a set-up cost, both of which are deterministic. For each product, the demand follows a Poisson process and the unit processing time is known. The relevant costs include the set-up cost, a cost per unit time while the machine is running, and linear costs for inventory and backlogging. The research objective is to find the production/inventory control policy that minimizes the expected total discounted cost over an infinite horizon.

Optimal control policies have been studied for decades for various single product production systems (see Sobel [12], Gavish and Graves [4], Srinivasan and Lee [13]). Chauny *et al.* [1] study the control problem for a single machine center multiproduct FMS (Flexible Manufacturing System) where no end-product inventory is allowed. They formulate the problem as a discrete semi-Markov decision process, which is solved by the Successive Approximation method in conjunction with interpolation techniques. Seidmann and Schweitzer [11] study the interaction of the flexible machine center and dedicated assembly lines that immediately follow. The linear cost charged

Manuscript received November 10, 1993; revised March 8, 1994.

The authors are with the Faculty of Management, McGill University, Montreal, P.Q., Canada H3A 1G5.
IEEE Log Number 9407000.

the time during which the assembly lines are idle is used as a surrogate measure of the throughput loss. This measure may, however, be difficult to justify when the goal of the production system is to produce to match the demand at the lowest cost rather than producing at the maximum rate. Graves [6] proposes a state-dependent heuristic solution procedure for the discrete-time, single machine multiproduct cycling problem. The simulation results show that the heuristic policies perform better than (s, S) or lot-sizing policies in most cases.

In this research, we allow both end-product inventory and backorders and develop a near-optimal production/inventory control policy for the production control problem. In Section II, we formulate the problem as a discrete semi-Markov decision process. In Section III, we develop a solution procedure which leads to a near-optimal policy and a procedure for generating its error bound. Computational test results are also presented to show the structure of the near-optimal policy and how its accuracy is affected by the system characteristics such as capacity utilization and set-up time. In Section IV, we conclude with a discussion on the future extension of this research.

II. DEVELOPMENT OF THE MODEL

Notations:

n	index for product, $1 \leq n \leq N$.
λ_n	demand rate of product n .
$d_n(t)$	cumulative demand of product n during time interval $[0, t]$; its distribution is Poisson with mean $\lambda_n t$.
τ_n	unit processing time of product n .
T_{ij}, W_{ij}	set-up time and cost when the machine switches production from product i to product j .
	$T_{n,n}, W_{n,n} = 0 \quad \forall n$.
c	production cost rate (\$/unit time).
g_n	inventory holding cost rate of product n (\$/unit time/unit).
g_{bn}	backlogging cost rate of product n (\$/unit time/unit).
$I(x_n)$	inventory/backlogging cost = $\begin{cases} g_n x_n & \text{if } x_n \geq 0 \\ -g_{bn} x_n & \text{if } x_n < 0. \end{cases}$
Q_n	limit on the end-product inventory of product n .
α	continuous discount rate, $\alpha > 0$.
$\text{Prob}(y = y_0)$	probability that y is equal to y_0 .
$E[y]$	expectation of random variable y .
$\exp(y)$	exponential function e^y .

The production/inventory control problem is formulated as a semi-Markov decision process (SMDP) (Denardo [3]) with the following characteristics.

State Space: The state of the system is comprised of the inventory state and machine set-up status. The inventory state space is defined as $X = \{x \in Z^N: x_n \leq Q_n, 1 \leq n \leq N\}$ with x_n representing the inventory (backlog if $x_n < 0$) level for product n . Let index k denote the machine status such that $k = n$ means that the machine is set up for product n . The state space of the system is thus defined as $S = \{s \in Z^{N+1}: s = (x, k)\}$.

Action Space: We assume that the production process is reviewed at those points in time when either the machine is idle and some demand arrives or when an item has just been processed and the machine becomes free. In other words, we assume that preemption of set-up or processing is not allowed. At any review point, one of $N+1$ actions is taken: let the machine stay idle ($a = 0$) and keep the latest set-up, or produce one of the N products ($a = n, 1 \leq n \leq N$).

The action space is thus defined as $A = \{a: a = 0, 1, \dots, N\}$. A stationary policy π consists of the set of actions taken depending on the observed state s .

Immediate Costs: Given any state of the system $s = (x, k)$, if action a ($a > 0$) is taken, the next decision point is the epoch at which one item of product a has just been processed. The discounted production cost during the time period $T_{ka} + \tau_a$ (note that $W_{ka} = T_{ka} = 0$ if $k = a$), $PC(k, a)$, includes set-up and variable costs and is given by

$$\begin{aligned} PC(k, a) &= W_{ka} + \int_{T_{ka}}^{T_{ka} + \tau_a} \exp(-\alpha \cdot t) \cdot c \cdot dt \\ &= W_{ka} + c \cdot \exp(-\alpha \cdot T_{ka}) \cdot \frac{1 - \exp(-\alpha \cdot \tau_a)}{\alpha}. \end{aligned}$$

The expected discounted inventory cost for product n during the same time period $T_{ka} + \tau_a$, $R_n(x_n, k, a)$, is given by

$$R_n(x_n, k, a) = \begin{cases} u_n(k, a) - b_n(k, a) \cdot x_n & \text{if } x_n \leq 0 \\ (g_n + g_{bn}) \cdot \sum_{i=0}^{x_n} (x_n - i) \cdot F_n(i, k, a) \\ \quad + u_n(k, a) - b_n(k, a) \cdot x_n & \text{if } x_n > 0 \end{cases}$$

where

$$u_n(k, a) = g_{bn} \cdot \lambda_n \cdot \left(\frac{1 - \exp(-\alpha \cdot (T_{ka} + \tau_a))}{\alpha^2} - \frac{(T_{ka} + \tau_a) \exp(-\alpha \cdot (T_{ka} + \tau_a))}{\alpha} \right)$$

$$b_n(k, a) = g_{bn} \cdot \frac{1 - \exp(-\alpha \cdot (T_{ka} + \tau_a))}{\alpha}. \quad (1)$$

See Appendix A for details of $F_n(i, k, a)$ and $R_n(x_n, k, a)$.

If the decision is to let the machine stay idle ($a = 0$) at state $s = (x, k)$, the next decision point is the epoch at which the next demand arrives. The expected discounted inventory cost for any product n until the next demand arrival is given by

$$\begin{aligned} R_n(x_n, k, a = 0) &= E_\theta \left[\int_0^\theta \exp(-\alpha \cdot t) \cdot I(x_n) \cdot dt \right] \\ &= E_\theta \left[\frac{1 - \exp(-\alpha \theta)}{\alpha} \right] \cdot I(x_n) \end{aligned}$$

where θ is the random variable representing the time between demand arrivals. The probability distribution of θ is exponential with parameter $\Lambda = \sum_{n=1}^N \lambda_n$. It follows that

$$R_n(x_n, k, a = 0) = \frac{I(x_n)}{\Lambda + \alpha} = \begin{cases} \frac{g_n}{\Lambda + \alpha} \cdot x_n & \text{if } x_n \geq 0 \\ -\frac{g_{bn}}{\Lambda + \alpha} \cdot x_n & \text{if } x_n < 0. \end{cases} \quad (2)$$

Therefore, the immediate cost for any action at state s , $C(s, a)$, can be expressed as

$$C(s, a) = PC(k, a) + \sum_{n=1}^N R_n(x_n, k, a) \quad (3)$$

where $PC(k, a = 0)$ is understood to be zero.

The dynamic programming recursion for minimizing the expected total discounted cost over an infinite horizon is as follows

$$\begin{aligned} V(s) &= \min \left\{ \left[C(s, a = 0) + \rho(k, a = 0) \right. \right. \\ &\quad \left. \left. + \sum_{n=1}^N \frac{\lambda_n}{\Lambda} \cdot V(x - e_n, k) \right]; \right. \\ &\quad \min_{a > 0} \left[C(s, a) + \rho(k, a) \cdot \sum_{j_1=0}^{\infty} P_1(j_1 | k, a) \cdots \right. \\ &\quad \left. \left. \cdot \sum_{j_N=0}^{\infty} P_N(j_N | k, a) \cdot V(x + e_a - j, a) \right] \right\} \quad (4) \end{aligned}$$

where

$$\rho(k, a) = \begin{cases} E_\theta[\exp(-a \cdot \theta)] = \frac{\lambda}{\lambda + a} & \text{if } a = 0 \\ \exp(-a \cdot (T_{k,a} + \tau_a)) & \text{if } a > 0 \end{cases}$$

$$e_a = \text{unit vector along } a^{\text{th}} \text{ coordinate}, \quad 1 \leq a \leq N$$

$$P_n(j_n | k, a > 0) = \text{Prob}(d_n(T_{k,a} + \tau_a) = j_n), \quad 1 \leq n \leq N$$

$$j = \text{vector}[j_n, 1 \leq n \leq N].$$

The first set of square brackets in (4) contains the cost for letting the machine stay idle until the next demand arrival. The second set of brackets contains the cost for producing one item of product a .

III. SOLUTION METHOD

Note that the model in (4) is defined on an infinite inventory state space which leads to unbounded costs. The general unbounded stochastic dynamic programming problems have been studied by many authors (see Whitt [15], Wessels [14], Lippman [10]). The major results have been summarized and extended by Van Nunen and Wessels [16]. In the context of our model in (4), the key results can be described as follows. Let μ be a positive function on S satisfying Assumption 2.3 in Van Nunen and Wessels [16]. (It could be shown that $\mu(s) = (\prod_{n=1}^N |x_n|)^{-1}$ or $\mu(s) = (\sum_{n=1}^N |x_n|)^{-1}$ satisfy the above assumption for our problem.) Denote by \bar{U} the class of real-valued functions on S with the property $\|v\| = \sup_{s \in S} |v(s)| \cdot \mu(s) < \infty$. Define on \bar{U} operators H for policy π and U^* , respectively, by

$$H_\pi v(s) = C(s, \pi(s)) + G(s, \pi(s), v)$$

where

$$G(s, \pi(s), v) = \begin{cases} \rho(k, a=0) \cdot \sum_{n=1}^N \frac{\lambda_n}{\lambda} \cdot v(x - e_n, k) & \text{if } \pi(s) = 0 \\ \rho(k, \pi(s)) \cdot \sum_{j_1=0}^{\infty} P_1(j_1 | k, \pi(s)) \cdots & \\ \sum_{j_N=0}^{\infty} P_N(j_N | k, \pi(s)) & \\ \cdot v(x + e_{\pi(s)} - j, \pi(s)) & \text{if } \pi(s) > 0 \end{cases} \quad (5)$$

and

$$U^* v(s) = \min \left\{ \left[C(s, a=0) + \rho(k, a=0) \cdot \sum_{n=1}^N \frac{\lambda_n}{\lambda} \cdot v(x - e_n, k) \right]; \right. \\ \left. \min_{a > 0} \left[C(s, a) + \rho(k, a) \cdot \sum_{j_1=0}^{\infty} P_1(j_1 | k, a) \cdots \right. \right. \\ \left. \left. \cdot \sum_{j_N=0}^{\infty} P_N(j_N | k, a) \cdot v(x + e_a - j, a) \right] \right\}. \quad (6)$$

Wessels [14] has shown that successive applications of U^* (respectively, H_π) converge in norm to V^* , the value function following the optimal policy (respectively, to V_π , the value function following policy π).

Due to the infinite inventory state space, however, the Successive Approximations scheme cannot be applied directly to our model. In Subsection A, a near-optimal policy is obtained by solving a problem defined on a truncated inventory state space. The challenge of our problem lies with the development of an error bound for using the near-optimal policy. A general approach proposed by Lippman [10] and Van Nunen and Wessels [16] for unbounded stochastic programming problems results in an extremely large error bound when applied to our policy. In Subsection B, we derive an alternative approach that yields a much tighter error bound on the same policy.

We also discuss the reasons for the improvement. In Subsection C we show the computational test results of the error bounds and the structure of our policies based on problems involving two products.

A. Approximation via a Finite-State Model

We first define a function that maps the infinite inventory state space onto a finite one. A policy is obtained by applying the Successive Approximations algorithm to a model defined on the finite inventory state space. This policy is then extended to construct a near-optimal policy for the original model.

Construct a finite inventory state space $\bar{X} = \{\bar{x} \in Z^N: L_n \leq \bar{x}_n \leq B_n, 1 \leq n \leq N\}$ through a mapping $p: X \mapsto \bar{X}$ defined by

$$\begin{cases} p(x_n) = x_n & \text{if } x_n \geq L_n \\ p(x_n) = L_n & \text{if } x_n < L_n \end{cases} \quad (7)$$

where $L = [L_n, 1 \leq n \leq N]$ can be interpreted as a truncation state in the original inventory state X . We restrict L_n to be less than zero for all n in order to allow for both positive and negative (backorders) inventories in \bar{X} . Define $\bar{S} = \{\bar{s} \in Z^{N+1}, \bar{s} = (\bar{x}, k)\}$. Clearly \bar{S} is a subset of S . Let \bar{U} be the class of bounded functions $\bar{U}(\cdot): \bar{S} \mapsto \mathcal{R}$, and $\bar{\pi}$ be a policy on \bar{S} . On \bar{U} we define operators $\bar{H}_{\bar{\pi}}$ for any policy $\bar{\pi}$, and \bar{U}^* , respectively, by

$$\bar{H}_{\bar{\pi}} \bar{v}(\bar{s}) = H_{\bar{\pi}} v_p(\bar{s}) \quad (8)$$

and

$$\bar{U}^* \bar{v}(\bar{s}) = U^* v_p(\bar{s}) \quad (9)$$

where $v_p(s)$ is a function in \bar{U} defined as $v_p(s) = \bar{v}(p(s), k)$ and π is a policy on S defined as $\pi(s) = \bar{\pi}(p(s), k)$.

Starting with an arbitrary function \bar{v}_0 in \bar{U} , iterations based on (9) are performed m times. Denote by $\bar{v}^{(m)}$ and $\bar{\pi}^{(m)}$, respectively, the value function and the policy determined by the last iteration. Based on the value function $\bar{v}^{(m)}$, two additional operations are performed as follows: $\bar{v}^{(m+1)} = \bar{H}_{\bar{\pi}^{(m)}} \bar{v}^{(m)}$ and $\bar{\pi}^{(m+1)} = \bar{U}^* \bar{v}^{(m)}$. Let $\delta^* = \max_{\bar{s} \in \bar{S}} |\bar{v}^{(m+1)}(\bar{s}) - \bar{v}^{(m)}(\bar{s})|$ and $\delta_* = \max_{\bar{s} \in \bar{S}} |\bar{v}^{(m)}(\bar{s}) - \bar{v}^{(m-1)}(\bar{s})|$. According to the contraction mapping theory (Denardo [3]), both δ^* and δ_* converge to zero as the number of iterations m goes to infinity.

In the next subsection, we derive a bound on the error, $V(s) - V_\pi(s)$, due to using policy π ($\pi(s) = \bar{\pi}(p(s), k)$), for the original production/inventory control problem defined in (4). We will also discuss how the error bound is affected by δ^* and δ_* .

B. Error Bound on the Near-Optimal Policy

Define a row vector $B(s) = [b_1(k, \pi(s)), \dots, b_N(k, \pi(s))]$, where $b_n(k, a > 0)$ is the immediate expected backlogging ($x_n < 0$) cost per unit defined in (1) and $b_n(k, a = 0) = g_{bn}/(\lambda + a)$ as indicated in (2).

Let $y^+ = \max\{y, 0\}$ and $(L - x)^+$ represent the column vector $[(L_n - x_n)^+, 1 \leq n \leq N]$. Define a sequence of value functions $u^{(i)}$ on S as follows:

$$u^{(1)}(s) = B(s) \cdot (L - x)^+ \quad (10)$$

$$u^{(i+1)}(s) = G(s, \pi(s), u^{(i)}), \quad i \geq 1 \quad (11)$$

where $G(s, \pi(s), v)$ in (11) is defined in (5).

Theorem 1. Let $\rho = \max_{k,a} \{\rho(k, a)\}$ and define a function $v_p^{(i)}(s)$ in \bar{U} as $v_p^{(i)}(s) = \bar{v}^{(i)}(p(s), k)$, then

$$V_\pi(s) \leq v_p^{(i)}(s) + \frac{\delta^*}{1 - \rho} + \sum_{i=1}^i u^{(i)}(s). \quad (12)$$

Proof Since $\pi(s) = \bar{\pi}(p(i) | k)$ by the definition of policy π immediate inventory cost $R_i(i, k | \pi(s))$ and the immediate cost $C(s | \pi(s))$ (both defined in Section II) can be rewritten as

$$l(i, k | \pi(s)) = R_n(p(i, k | \pi(s))) + b(k | \pi(s)) (L - i)^+$$

$$C(s | \pi(s)) = PC(k | \pi(s)) + \sum_1^N R(i, k | \pi(s)) \\ = C(p(i) | k | \pi(s)) + u^{(1)}(s)$$

$$H_{\pi} i_j'(s) = C(s | \pi(s)) + G(s | \tau(s) | i_j) \\ = i_j^{-1}(p(i) | k) + u^{(1)}(s) \\ \leq i_j(s) + \delta^* + u^{(1)}(s)$$

$$H_{\pi} i_j'(s) \leq H_{\pi}(i_j + \delta^* + u^{(1)}) \\ = C(s | \pi(s)) + G(s | \tau(s) | i_j + \delta^* + u^{(1)}) \\ \leq i_j(s) + (1 + \rho) \delta^* + u^{(1)}(s) + u^{(1)}(s)$$

Since $H_{\pi} i_j \rightarrow V_{\pi}$ we obtain

$$V_{\pi}(s) \leq i_j(s) + \frac{\delta^*}{1 - \rho} + \sum_1^N u^{(1)}(s) \quad \square$$

Theorem 2

$$V(s) = E^{\infty} i_j(s) \geq i_j(s) - \frac{\delta_*}{1 - \rho} \quad (13)$$

Theorem 2 is proved in a way similar to Theorem 1. Theorems 1 and 2 lead to the following result

$$i_j(s) - \frac{\delta_*}{1 - \rho} \leq V(s) \leq V_{\pi}(s) \leq i_j(s) + \frac{\delta^*}{1 - \rho} + \sum_1^N u^{(1)}(s)$$

Computing an Upper Bound on $\sum_1^N u^{(1)}(s)$ The exact value of $\sum_1^N u^{(1)}(s)$ is difficult to compute since $u^{(1)}(s)$ itself may involve infinite arithmetic operations (when $\pi(s) \sim 0$). In what follows we provide a procedure for finding its upper bound.

First a crude upper bound on $u^{(1)}(s)$ $\bar{u}^{(1)}(s)$ can be easily derived and is given as follows

$$\bar{u}^{(1)}(s) = \rho^{-1} B^* [(L - i)^+ + (i - 1) - V^*] \quad \forall s \in S, i \geq 1$$

where the row vector $B^* = [\max_k \{b(k | a)\} \quad 1 \leq n \leq N]$ and the column vector $V^* = [\max \{V / \Lambda \mid \max_k \{F[d(\tau + I_k)]\} \quad 1 \leq n \leq N\}]$. Note that $F[d(\tau + I_k)] = V(\tau + I_k)$. Given an integer M we obtain

$$\sum_1^N u^{(1)}(s) \leq \sum_1^M u^{(1)}(s) + \sum_{M+1}^N \bar{u}^{(1)}(s) \\ = \sum_1^M u^{(1)}(s) + \rho^M B^* \\ \left[\frac{(L - i)^+}{1 - \rho} + \frac{M(1 - \rho) + \rho}{(1 - \rho)^2} - V^* \right]$$

Clearly when M is large the right-hand side of the above inequality is close to $\sum_1^N u^{(1)}(s)$

Second define a sequence of value functions $u^{(n)}(s)$ as follows

$$\bar{u}^{(1)}(s) = u^{(1)}(s) \quad \forall s \in S$$

$$u^{(n)}(s) = \bar{u}^{(n)}(s) \quad \text{if } i < I \quad \text{for some } n \leq i - 1$$

$$u^{(n+1)}(s) = G(s | \pi(s) | u^{(n)}) \quad \text{if } i \geq I \quad \text{for all } n \leq i - 1$$

It can be shown by induction that $u^{(i)}(s) \leq u^{(i-1)}(s)$. Appendix B shows that $u^{(i)}(s)$ can be computed within finite arithmetic. Let

$$\delta(s) = \sum_1^M u^{(i)}(s) + \rho^M B^* \\ \left[\frac{(L - i)^+}{1 - \rho} + \frac{M(1 - \rho) + \rho}{(1 - \rho)^2} - V^* \right]$$

The bound on the error due to using near optimal policy $\tau | \pi(s) - V(s)$ is thus given by

$$\epsilon(s) = \frac{\delta_* + \delta}{1 - \rho} + \delta(s) \quad (14)$$

Remark 1 The error bound in (14) has three components: δ_* , δ^* and δ where δ_* and δ^* measure the error due to the Successive Approximations on the truncated model. Given any truncated inventory state space δ_* and δ^* can be made arbitrarily small by performing a large number of iterations based on (9). $\delta(s)$ on the other hand measures the error due to truncation and it can only be made small by reducing L . The last point can be demonstrated by the numerical results reported in the next subsection. Consequently the quality of the near optimal solution is mainly determined by the selection of the truncation state L .

Remark 2 A general approach for deriving an error bound for unbounded stochastic dynamic programming problems can be found in Lippman [10] and Van Nunen and Wessels [16]. The key element to the general approach is the norm $\|H_{\pi} i_j - i_j\|$ defined as $\max_{s \in S} \{|H_{\pi} i_j(s) - i_j'(s)| \mid \mu(s)\}$. We have tried two forms of μ : $\mu(s) = (\prod_{i=1}^N |i|)^{-1}$ and $\mu(s) = (\sum_{i=1}^N |i|)^{-1}$ but both yield extremely large error bounds (100-1000% of the lower bound in all instances tested). More critically there is no indication that the error bound would approach zero when the truncation state I decreases to $-\infty$. In contrast the results based on our approach show (in the next subsection) that the error for using policy π is bounded by a rather small value when the initial inventory state is reasonably far above the truncation state. The improvement in our view comes from the fact that our approach explicitly uses the Poisson probability distribution (for computing $u^{(1)}$ in δ) which guarantees that the probability of reaching the inventory state below the current one decreases exponentially. As a result the impact of the value function in inventory states below the truncation state which is not captured by the Successive Approximations procedure on the truncated state space is not significant.

C. Experimental Results Based on Two Product Problems

We now consider a manufacturing system that produces two homogeneous products. Four experiments (shown in Table I) are conducted to examine the impact of set-up time and capacity utilization denoted by τ ($\tau = 2 - \lambda \tau$, for our experiments), on the error bound of the near-optimal policies.

For each problem a near-optimal policy is computed based on the iteration procedure (9) with the following stopping criterion

$$\max_{s \in S} \left\{ \frac{\delta_* + \delta}{1 - \rho} \right\} \leq 0.001$$

TABLE I
FOUR EXPERIMENTS

Parameters	Experiments			
	Problem 1	Problem 2	Problem 3	Problem 4
$T_{12} = T_{21} = T$	$T = 1.0$	1.0	2.0	2.0
$\lambda_1 = \lambda_2 = \lambda$	0.25	0.40	0.25	0.40
$\tau_1 = \tau_2 = \tau$	1.0	1.0	1.0	1.0
$r = 2\lambda\tau$	0.50	0.80	0.50	0.80
$c_1 = c_2$	\$3000	\$3000	\$3000	\$3000
$W_1 = W_2$	\$50	\$50	\$50	\$50
$g_1 = g_2$	\$100	\$100	\$100	\$100
$g_{b1} = g_{b2}$	\$500	\$500	\$500	\$500
$Q_1 = Q_2$	50	50	50	50
$L_1 = L_2$	-50	-50	-50	-50
α	0.025	0.025	0.025	0.025

The quantity on the left-hand side of the above inequality represents the relative error bound due to the imperfect convergence of the Successive Approximations procedure for generating the near-optimal policy.

We then compute the maximum of the relative error due to truncation, defined as $\delta(s)/\bar{v}'''(s)$, over five different inventory state regions. The results are presented in Table II. The following two observations are of particular interest.

1) In each experiment, the error due to truncation tends to increase as the initial inventory state moves closer to the truncation state L , and becomes extremely large when the initial inventory level for both products is below -35 [with $L = (-50, -50)$]. This means that the near-optimal policy can be a good approximation of the true optimal one only when the initial inventory state is reasonably far above the truncation state L . Therefore, for a given initial inventory state, the near-optimal policy can be improved by reducing L . This obviously means increased computing time required for generating the near-optimal policy since the iterations based on (9) have to be performed over a larger inventory state space.

2) Given the same region of initial inventory states, the error due to truncation tends to increase as the capacity utilization and/or set-up time increase. Intuitively, this is because the probability of the inventory state moving beyond the truncation state in the future transitions increases as the machine becomes busier. This suggests that the truncation state L required for generating good near-optimal policies is determined by the machine capacity utilization and/or set-up time; the higher the machine capacity utilization and/or set-up time, the lower the truncation state has to be (which means more computing time required as a result of observation 1).

Fig. 1 depicts the structure of the near-optimal policy for the problem with $r = 0.5$ and $T = 1.0$. The two coordinates represent the inventory positions for products 1 and 2, respectively. The solid (dashed) lines delimit the actions to be taken when the machine is set up for product 1 (product 2). These lines (also called *switching curves*), create four regions, and in any one region the action is well defined. Thus, in region I (respectively, region III), one should produce product 1 (product 2, respectively), irrespective of the current machine status. In region II, do not produce anything, but keep current machine set-up. In region IV, continue to produce current product (whether 1 or 2). The shape of the policy clearly indicates the monotonicity property: given the machine status and inventory level

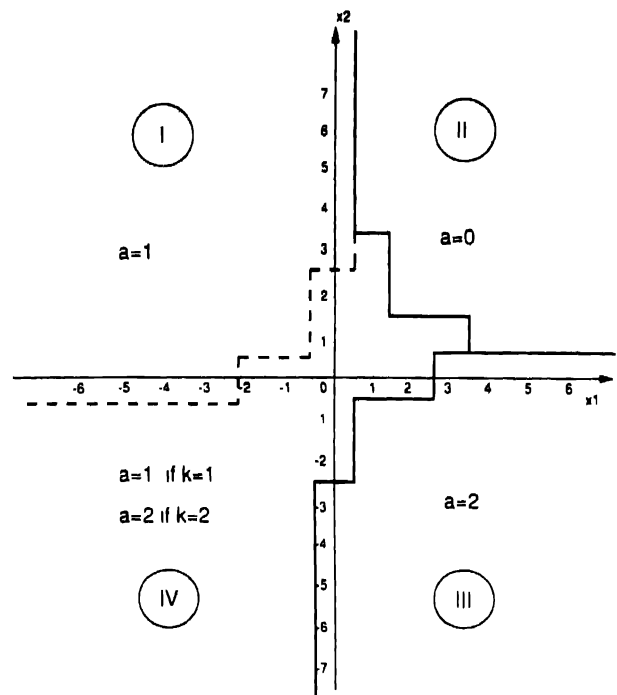


Fig. 1. The near-optimal policy

for one product, there is a unique threshold level for the other product to switch the decision, either to stop production or start producing the other one. It also shows that the near-optimal policy is not of the classical (s, Q) or (s, S) types, since the points at which production switches depend on *both* inventory levels and the quantities produced depend on the quantities of the realized demands for *both* products during the production run.

IV. CONCLUSIONS

In this note, we have developed a procedure that generates near-optimal policies for the single machine multiproduct production/inventory control problem by solving a problem defined on a truncated inventory state space. The experiment test results have shown that the error for using a near-optimal policy is bounded by a small value if the initial inventory state is reasonably far above the truncation state. We have also found that policies based on predetermined structures such as lot-sizing and (s, S) policies can be very different from the optimal policy and therefore perform poorly in practice. For a problem that involves many products, however, our procedure cannot be both efficient and accurate due to the "curse of dimensionality." For a ten-product problem, for instance, even if we preserve only five inventory states for each product in the truncated space, the total number of inventory states becomes $5^{10} \times 10 = 9.76 \times 10^7$.

We are in the process of developing a decomposition/aggregation solution procedure for problems with many products. For an N -product problem, the procedure decomposes the problem into N subproblems, each involving a single product and a composite product that represents the remaining $N - 1$ products. Based on the solutions of the two-product subproblems using the method developed in this note, the procedure constructs a separate policy for each individual product such that policies for different products do not conflict at any state of the system. Preliminary results have shown that this procedure leads to very good policies.

TABLE II
RELATIVE ERROR BOUNDS OF THE NEAR-OPTIMAL POLICIES DUE TO TRUNCATION

Parameters		Inventory State Regions				
		$x_1 \geq -15$	$x_1 \geq -20$	$x_1 \geq -25$	$x_1 \geq -35$	$x_1 \geq -50$
		$x_2 \geq -15$	$x_2 \geq -20$	$x_2 \geq -25$	$x_2 \geq -35$	$x_2 \geq -50$
$r = 0.5$	$T = 1.0$	1.56×10^{-13}	3.37×10^{-9}	9.76×10^{-6}	8.89×10^{-2}	1.61
	$T = 2.0$	4.56×10^{-13}	9.46×10^{-9}	2.61×10^{-5}	0.21	3.57
$r = 0.8$	$T = 1.0$	4.17×10^{-5}	4.07×10^{-3}	9.64×10^{-3}	0.74	2.71
	$T = 2.0$	1.38×10^{-4}	1.16×10^{-2}	0.24	1.65	8.66

APPENDIX A

THE INVENTORY COST DURING UNIT PROCESSING TIME

If $x_n < 0$,

$$\begin{aligned} R_n(x_n, k, a) &= \int_0^{T_{ka} + \tau_a} \exp(-\alpha \cdot t) \cdot g_{bn} \cdot E[d_n(t) - x_n] \cdot dt \\ &= \int_0^{T_{ka} + \tau_a} \exp(-\alpha \cdot t) \cdot g_{bn} \cdot (\lambda_n \cdot t - x_n) \cdot dt \\ &= a_n(k, a) - b_n(k, a) \cdot x_n. \end{aligned}$$

If $x_n \geq 0$,

$$\begin{aligned} R(x_n, k, a) &= \int_0^{T_{ka} + \tau_a} \exp(-\alpha \cdot t) \cdot \{g_n \cdot E[(x_n - d_n(t))^+] \\ &\quad + g_{bn} \cdot E[(x_n - d_n(t))^-]\} \cdot dt \\ &= \int_0^{T_{ka} + \tau_a} \exp(-\alpha \cdot t) \\ &\quad \cdot \left[g_n \cdot \sum_{i=0}^{x_n} (x_n - i) \cdot \text{Prob}(d_n(t) = i) \right. \\ &\quad \left. + g_{bn} \cdot \sum_{i=x_n+1}^{\infty} (i - x_n) \cdot \text{Prob}(d_n(t) = i) \right] \cdot dt \\ &= \int_0^{T_{ka} + \tau_a} \exp(-\alpha \cdot t) \\ &\quad \cdot \left[(g_n + g_{bn}) \cdot \sum_{i=0}^{x_n} (x_n - i) \frac{(\lambda_n \cdot t)^i \exp(-\lambda_n \cdot t)}{i!} \right. \\ &\quad \left. + \int_0^{T_{ka} + \tau_a} \exp(-\alpha \cdot t) \cdot g_{bn} \cdot (\lambda_n \cdot t - x_n) \cdot dt \right] \cdot dt. \end{aligned}$$

The second term is exactly $a_n(k, a) - b_n(k, a) \cdot x_n$. The first term becomes

$$\begin{aligned} (g_n + g_{bn}) \cdot \sum_{i=0}^{x_n} (x_n - i) \cdot \int_0^{T_{ka} + \tau_a} \exp(-(\lambda_n + \alpha) \cdot t) \\ \cdot \frac{(\lambda_n \cdot t)^i}{i!} \cdot dt = (g_n + g_{bn}) \cdot \sum_{i=0}^{x_n} (x_n - i) \cdot F_n(i, k, a) \end{aligned}$$

where $F_n(i, k, a)$ can be computed using the recursive relationship

$$\begin{aligned} F_n(i, k, a) &= \int_0^{T_{ka} + \tau_a} \exp(-(\lambda_n + \alpha) \cdot t) \cdot \frac{(\lambda_n \cdot t)^i}{i!} \cdot dt \\ &= \frac{(\lambda_n \cdot (T_{ka} + \tau_a))^i \cdot \exp(-(\lambda_n + \alpha) \cdot (T_{ka} + \tau_a))}{-(\lambda_n + \alpha) \cdot i!} \\ &\quad + \frac{\lambda_n}{\lambda_n + \alpha} \cdot F_n(i-1, k, a) \end{aligned}$$

with $F_n(0, k, a) = \int_0^{T_{ka} + \tau_a} \exp(-(\lambda_n + \alpha) \cdot t) \cdot dt = [1 - \exp(-(\lambda_n + \alpha) \cdot (T_{ka} + \tau_a))]/(\lambda_n + \alpha)$.

APPENDIX B

COMPUTATION OF THE FUNCTIONS $\hat{u}^{(i)}$

The function $\hat{u}^{(i)}(s)$ is easy to compute at state s where $\pi(s) = 0$. The following is the procedure for computing $\hat{u}^{(i)}(s)$ at state s where $\pi(s) > 0$:

$$\begin{aligned} \hat{u}^{(i+1)}(s) &= G(s, \pi(s), \hat{u}^{(i)}) \\ &= \rho(k, \pi(s)) \cdot \sum_{j_1=0}^{x_1 + e_{\pi(s), 1} - L_1} P_1(j_1 | k, \pi(s)) \cdots \\ &\quad \cdot \sum_{j_N=0}^{x_N + e_{\pi(s), N} - L_N} P_N(j_N | k, \pi(s)) \\ &\quad \cdot \hat{u}^{(i)}(x + e_{\pi(s)} - j, \pi(s)) + \rho(k, \pi(s)) \cdot \rho^{i-1} \cdot B^* \\ &\quad \cdot \left\{ (i-1) \cdot \left[1 - \prod_{n=1}^N \sum_{j=0}^{x_n + e_{\pi(s), n} - L_n} P_n(j | k, \pi(s)) \right] \right. \\ &\quad \left. \cdot A^* + D(s) \right\} \end{aligned}$$

where the column vector $D(s)$ is given by

$$\begin{aligned} D(s) &= \left[\lambda_n \cdot (\tau_{\pi(s)} + T_{k, \pi(s)}) \right. \\ &\quad \left. - \sum_{j=0}^{x_n + e_{\pi(s), n} - L_n} P_n(j | k, \pi(s)) \cdot j, 1 \leq n \leq N \right] \end{aligned}$$

and $e_{\pi(s), n}$ represents the n^{th} entry in vector $e_{\pi(s)}$, $1 \leq n \leq N$.

REFERENCES

- [1] F. Chauny, A. Hauric, P. L'Ecuyer, and R. Loulou, "Dynamic programming solution to the stochastic multiple lot dispatching in an FMS," in *Proc. 1st Rensselaer Int. Conf. Comput. Integrated Manuf.*, 1988, pp. 238-243.
- [2] E. V. Denardo and C. S. Tang, "Linear control of a Markov production system," *Oper. Res.*, vol. 40, pp. 259-278, 1992.
- [3] E. V. Denardo, "Contraction mappings in the theory underlying dynamic programming," *SIAM Rev.*, vol. 9, pp. 165-177, Apr. 1967.
- [4] B. Graves and C. S. Graves, "A one-product production/inventory problem under continuous review policy," *Oper. Res.*, vol. 28, pp. 1228-1236, Sept.-Oct. 1980.
- [5] C. S. Graves, "A tractical planning model for a job shop," *Oper. Res.*, vol. 34, pp. 522-533, 1986.
- [6] —, "The multi-product production cycling problem," *AIIE Trans.*, vol. 12, pp. 233-240, Sept. 1980.
- [7] —, "Safety stocks in manufacturing systems," *J. Manuf. Oper. Manage.*, vol. 1, pp. 67-101, Spring 1988.
- [8] U. S. Karmarkar, "Kanban systems," Working Paper Series No. QM8325, Grad. School Manage., Univ. Rochester, 1986.
- [9] U. S. Karmarkar, S. Kekre, and S. Kekre, "Multi-item lot sizing and manufacturing leadtimes," Working Paper Series No. QM8325, Grad. School Manage., Univ. Rochester, 1983.

- [10] S. A. Lippman, "Semi-Markov decision processes with unbounded rewards," *Manage. Sci.*, vol. 19, pp. 717-737, July 1975.
- [11] A. Seidmann and P. A. Schweitzer, "Part selection policy for a flexible manufacturing cell feeding several production lines," *IEE Trans.*, vol. 16, pp. 355-362, Aug. 1984.
- [12] M. J. Sobel, "Optimal average-cost policy for a queue with start-up and shut-down costs," *Oper. Res.*, vol. 17, pp. 145-162, Jan.-Feb. 1969.
- [13] M. M. Srinivasan and H. Lee, "Random review production/inventory systems with compound Poisson demands and arbitrary processing times," *Manage. Sci.*, vol. 37, pp. 813-833, July 1991.
- [14] J. Wessels, "Markov programming by successive approximations with respect to weighted supremum norms," *J. Math. Anal. Appl.*, vol. 58, pp. 326-335, 1977.
- [15] W. Whitt, "Approximations of dynamic programs II," *Math. Oper. Res.*, vol. 3, pp. 179-185, May 1979.
- [16] J. A. E. Van Nunen and J. Wessels, "Markov decision processes with unbounded rewards," in *Markov Decision Theory*, H. Tijms and J. Wessels, Eds. Mathematical Center Tract No. 93, Amsterdam, 1977, pp. 1-24.
- [17] P. H. Zipkin, "Models for design and control of stochastic, multi-item batch production systems," *Oper. Res.*, vol. 34, pp. 91-104, Jan.-Feb. 1986.

The Partial Model Matching or Partial Disturbance Rejection Problem: Geometric and Structural Solutions

Michel Malabre and Juan Carlos Martinez Garcia

Abstract—The Partial Model Matching Problem (or equivalently, the Partial Disturbance Rejection Problem) is revisited here. It has been initially introduced in [2] and amounts to matching the first k Markov parameters of the plant with those of the model. This obviously finds all its interest when no solution exists to the Exact Problem. We give both geometric and structural solutions to these problems. The geometric solution is in terms of the steps of the well known supremal output nulling controlled invariant subspace algorithm. The structural conditions amount to comparing some list of integers, namely some orders of zeros at infinity.

I. NOTATION AND BASIC CONCEPTS

Throughout the paper we shall essentially follow the notational conventions of [11]. Script capital $(\mathcal{X}, \mathcal{Y}, \dots)$ denote finite-dimensional vector spaces over the field of real numbers \mathbb{R} , and $\dim(\mathcal{X}), \dim(\mathcal{Y}), \dots$ denote their dimensions. The notation $\mathcal{X} \simeq \mathcal{Y}$ means $\dim(\mathcal{X}) = \dim(\mathcal{Y})$. If $\mathcal{V} \subset \mathcal{X}$, then \mathcal{X}/\mathcal{V} denotes the quotient space \mathcal{X} modulo \mathcal{V} .

Italic capitals (A, B, \dots) denote interchangeably linear maps and their matrix representations in particular bases. The image of a map B is written as $\text{Im}B$ and its kernel as $\text{Ker}B$. The identity map on a n -dimensional space is denoted by I_n .

It is assumed here that the reader is familiarized with the concepts of (A, B) and (C, A) -invariance [11].

Manuscript received November 24, 1993; revised March 11, 1994. This work was supported in part by the National Council of Science and Technology of Mexico, the Advanced Studies and Research Center of the IPN of Mexico, and ESPRIT Basic Research Project, 8924 (SESDIP).

M. Malabre is with the Laboratoire d'Automatique de Nantes CNRS URA 823 E.C.N., 1 rue de la Noë, 44072 Nantes Cédex 03, France.

J. C. Martinez Garcia was with the Laboratoire d'Automatique de Nantes CNRS URA 823 E.C.N., 1 rue de la Noë, 44072 Nantes Cédex 03, France and is now with CINVESTAV, Mexico.

IEEE Log Number 9407218.

We shall consider linear time-invariant systems described by

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), & t \geq 0 \\ y(t) = Cx(t), & t \geq 0 \end{cases} \quad (1)$$

with $A: \mathcal{X} \rightarrow \mathcal{X}$, $B: \mathcal{U} \rightarrow \mathcal{X}$, $C: \mathcal{X} \rightarrow \mathcal{Y}$ ($\dim(\mathcal{X}) = n$, $\dim(\mathcal{U}) = m$, $\dim(\mathcal{Y}) = p$). This system will be noted by (A, B, C) .

Let us denote:

- $\mathcal{B} = \text{Im}B$.
- $\mathcal{K} = \text{Ker}C$.
- $\mathcal{J}(A, B; \mathcal{K})$ the family of (A, B) -invariant subspaces contained in \mathcal{K} , also called output nulling controlled invariant subspaces.
- $\mathcal{C}(C, A; \mathcal{B})$ the family of (C, A) -invariant subspaces containing \mathcal{B} .
- \mathcal{V}^* the supremal element of $\mathcal{J}(A, B; \mathcal{K})$. \mathcal{V}^* is equal to the limit of the Invariant Subspace Algorithm (ISA) for (A, B, C)

$$\begin{cases} \mathcal{V}^0 := \mathcal{X} \\ \mathcal{V}^i := \mathcal{K} \cap A^{-1}(\mathcal{B} + \mathcal{V}^{i-1}), & i \geq 1. \end{cases} \quad (2)$$

- \mathcal{S}^* the infimal element of $\mathcal{C}(C, A; \mathcal{B})$. \mathcal{S}^* is equal to the limit of the Conditioned Invariant Subspace Algorithm (CISA)

$$\begin{cases} \mathcal{S}^0 := 0 \\ \mathcal{S}^i := \mathcal{B} + A(\mathcal{K} \cap \mathcal{S}^{i-1}), & i \geq 1. \end{cases} \quad (3)$$

- $\mathcal{F}(\mathcal{V}^*)$ the family of static state feedback maps such that $(A + BF)\mathcal{V}^* \subset \mathcal{V}^*$.

Infinite Zero Structure

Given any system (A, B, C) described by (1) or equivalently by its strictly proper $(p \times m)$ transfer function matrix

$$T(s) := C(sI_n - A)^{-1}B \quad (4)$$

its structure at infinity, since such systems have no poles at infinity, is described by the multiplicity orders of its zeros at infinity. From an algebraic point of view, this structure can be derived from the so-called Smith-McMillan Form at infinity of $T(s)$, say Λ_∞ , which is a canonical form under right and left biproper transformations (see for instance [10]). Indeed, there exist biproper matrices, $D_1(s)$ and $D_2(s)$, such that

$$D_1(s)T(s)D_2(s) = \Lambda_\infty = \begin{bmatrix} \Delta_\infty & 0 \\ 0 & 0 \end{bmatrix}$$

where $\Delta_\infty = \text{diag}\{s^{-n_1}, s^{-n_2}, \dots, s^{-n_r}\}$, $r := \text{rank}(T(s))$.

The nonincreasing list of integers $\{n_i\}$ is the list of the orders of the zeros at infinity of the system. From a geometric point of view, various equivalent definitions have been given for this structure. The original one, due to [1], is

$$n_i = \text{card}\{p_i \geq i\} \quad \forall i \in \{1, \dots, r\} \quad (5)$$

where card stands for cardinal (number of elements in the set) and with

$$p_i := \dim \left(\frac{\mathcal{V}^* + \mathcal{S}^{i-1}}{\mathcal{V}^* + \mathcal{S}^{i-1}} \right), \quad \forall i \geq 1. \quad (6)$$

Other geometric characterizations have been given in [5]. Indeed

$$\begin{aligned} p_i &= \dim \left(\frac{\mathcal{V}^* + \mathcal{S}^{i-1}}{\mathcal{V}^* + \mathcal{S}^{i-1}} \right) \\ &= \dim \left(\frac{\mathcal{B} \cap \mathcal{V}^{i-1}}{\mathcal{B} \cap \mathcal{V}^*} \right) \\ &= \dim \left(\frac{\mathcal{S}^* \cap \mathcal{V}^{i-1}}{\mathcal{S}^* \cap \mathcal{V}^*} \right), \quad \forall i \geq 1. \end{aligned} \quad (7)$$

The following lemma will help us in the sequel.

Lemma 1: Let $\mathcal{V}^i, i \geq 0$, denote the subspace obtained in the i th step of (ISA). There exists a (non unique) state feedback F_0 in $\mathcal{F}(\mathcal{V}^*)$ satisfying

$$\mathcal{V}^{i+1} = \mathcal{K} \cap A_{F_0}^{-1} \mathcal{V}^i, \forall i \geq 0 \quad (8)$$

where $A_{F_0} := (A + BF_0)$.

Remark 1: Due to the correspondence (5) between list $\{n_j\}$ and $\{p_i\}$, it is quite obvious that the list

$$\{p_1 - p_i\}, \forall i \in \{1, 2, \dots, n_1\} \quad (9)$$

is the list of the orders of the zeros at infinity of (A, B, C) which are strictly smaller than i .

II. PROBLEMS STATEMENT

The k th-Order Partial Model Matching Problem (PMMP(k)):

Given a linear time-invariant system described by (1) and with associated transfer function matrix described by (4), and given a linear time-invariant prespecified model described by

$$\begin{cases} \dot{x}_m(t) = A_m x_m(t) + B_m u_m(t), & t \geq 0 \\ y_m(t) = C_m x_m(t), & t \geq 0, \end{cases} \quad (10)$$

with the state $x_m \in \mathcal{X}_m \simeq \mathcal{R}^{n_m}$, the input $u_m \in \mathcal{U}_m \simeq \mathcal{R}^{m_m}$ and the output $y_m \in \mathcal{Y} \simeq \mathcal{R}^p$, and with associated transfer function matrix

$$T_m(s) = C_m(sI_{n_m} - A_m)^{-1} B_m \quad (11)$$

we can formulate the k th order Partial Model Matching Problem (initially appeared in [2]), as follows:

Definition 1: Let $k \in \mathcal{N}$ be given. The k th order Partial Model Matching Problem has a solution if and only if there exists a proper rational matrix $C(s)$ such that

$$T(s)C(s) - T_m(s) = s^{-(k+1)}P(s) \quad (12)$$

for some proper $P(s)$. The proper transfer function matrix $C(s)$ can be interpreted as a dynamic precompensator cascaded with the given plant. In other terms, we want to match the first k Markov parameters of the compensated plant with those of the model.

We give Definition 1 for strictly proper plant and model. This assumption is not restrictive: the main interest, for the exposition, is that the associated geometry is much simpler. Extension to the proper case is quite easy.

Note that, because of the assumption of strict properness, (12) is trivially satisfied for $k = 0$. We shall indeed consider in the sequel that $k \geq 1$.

In a similar way, we shall consider the following.

The k th-Order Partial Disturbance Rejection Problem (PDRP(k))

PDRP(k) is defined as follows:

Definition 2: Given a perturbed linear, time-invariant, system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + Ed(t), & t \geq 0 \\ y(t) = Cx(t), & t \geq 0 \end{cases} \quad (13)$$

that we shall denote by $(A, [B \ E], C)$, with the state $x \in \mathcal{X} \simeq \mathcal{R}^n$, the input $u \in \mathcal{U} \simeq \mathcal{R}^m$, the output $y \in \mathcal{Y} \simeq \mathcal{R}^p$ and the disturbance $d \in \mathcal{D} \simeq \mathcal{R}^q$ (of course some smoothness properties of $d(\cdot)$, like measurability, shall be implicitly assumed here), and an integer $k \geq 1$, find necessary and sufficient conditions for the existence of a state feedback map $F: \mathcal{X} \rightarrow \mathcal{U}$ such that

$$\begin{cases} CE = 0 \\ C(A + BF)E = 0 \\ \vdots \\ C(A + BF)^{k-1}E = 0. \end{cases} \quad (14)$$

When the control law includes some measurements of the disturbance, we have a modified version of PDRP(k), namely the following.

The k th-Order Partial Modified Disturbance Rejection Problem (PMDRP(k))

Definition 3: Given the perturbed linear, time-invariant, system described by (13) and an integer $k \geq 1$, find necessary and sufficient conditions for the existence of a static state feedback law $u(t) = Fx(t) + Gd(t)$, with $F: \mathcal{X} \rightarrow \mathcal{U}$ and $G: \mathcal{D} \rightarrow \mathcal{U}$, such that

$$\begin{cases} C(E + BG) = 0 \\ C(A + BF)(E + BG) = 0 \\ \vdots \\ C(A + BF)^{k-1}(E + BG) = 0. \end{cases} \quad (15)$$

We shall provide in the next sections geometric and structural solutions to these problems. We shall first consider here PDRP(k).

III. GEOMETRIC SOLVABILITY CONDITIONS

Let us now consider the geometric solvability conditions of both problems introduced in Definition 2 and Definition 3.

Theorem 1: Let $k \geq 1$ be given, there exists a static state feedback map F which solves the k th order Partial Disturbance Rejection Problem if and only if

$$\mathcal{E} \subset \mathcal{V}^k \quad (16)$$

where \mathcal{E} denotes $\text{Im}E$ and \mathcal{V}^k denotes the subspace obtained at the k th step of (ISA) for (A, B, C) .

Proof of Theorem 1: Let \mathcal{K} denotes the kernel of C . If F exists, which satisfies (14), obviously

$$\begin{aligned} \mathcal{E} &\subset \mathcal{K} \\ (A + BF)\mathcal{E} &\subset \mathcal{K} \\ &\vdots \\ (A + BF)^{k-1}\mathcal{E} &\subset \mathcal{K} \end{aligned}$$

thus

$$\mathcal{E} \subset \mathcal{K} \cap A_F^{-1}\mathcal{K} \cap A_F^{-2}\mathcal{K} \cdots \cap A_F^{-(k-1)}\mathcal{K} \quad (17)$$

where $A_F := (A + BF)$. Now, it follows from (17) and (2) that

$$\begin{aligned} \mathcal{K} \cap A_F^{-1}\mathcal{K} &\subset \mathcal{V}^2 := \mathcal{K} \cap A^{-1}(B + \mathcal{K}) \\ &= \mathcal{K} \cap A_F^{-1}(B + \mathcal{K}), \\ \mathcal{K} \cap A_F^{-1}\mathcal{K} \cap A_F^{-2}\mathcal{K} &= \mathcal{K} \cap A_F^{-1}(\mathcal{K} \cap A_F^{-1}\mathcal{K}) \\ &\subset \mathcal{K} \cap A_F^{-1}\mathcal{V}^2 \\ &\subset \mathcal{K} \cap A_F^{-1}(\mathcal{V}^2 + \mathcal{B}) \\ &= \mathcal{K} \cap A^{-1}(B + \mathcal{V}^2) = \mathcal{V}^3, \\ &\vdots \\ \mathcal{K} \cap A_F^{-1}\mathcal{K} \cap A_F^{-2}\mathcal{K} \cdots \cap A_F^{-(k-1)}\mathcal{K} &\subset \mathcal{K} \cap A_F^{-1}(\mathcal{V}^{k-1} + \mathcal{B}) \\ &= \mathcal{K} \cap A^{-1}(B + \mathcal{V}^{k-1}) = \mathcal{V}^k \end{aligned}$$

then $\mathcal{E} \subset \mathcal{V}^k$.

Conversely, from Lemma 1, there exists at least one map F_0 such that

$$\mathcal{V}^{i+1} = \mathcal{K} \cap A_{F_0}^{-1}\mathcal{V}^i, \forall i \geq 0$$

where $A_{F_0} := (A + BF_0)$. Hence

$$\begin{aligned} \mathcal{E} \subset \mathcal{V}^k &\iff \mathcal{E} \subset \mathcal{K} \cap A_{F_0}^{-1} \mathcal{V}^{k-1}, \\ &\iff \begin{cases} \mathcal{E} \subset \mathcal{K} \\ A_{F_0} \mathcal{E} \subset \mathcal{V}^{k-1} = \mathcal{K} \cap A_{F_0}^{-1} \mathcal{V}^{k-2}, \\ \mathcal{E} \subset \mathcal{K} \\ A_{F_0} \mathcal{E} \subset \mathcal{K} \\ A_{F_0}^2 \mathcal{E} \subset \mathcal{V}^{k-2}, \\ \vdots \\ \mathcal{E} \subset \mathcal{K} \\ A_{F_0} \mathcal{E} \subset \mathcal{K} \\ \vdots \\ A_{F_0}^{k-1} \mathcal{E} \subset \mathcal{K} \end{cases} \\ &\iff \begin{cases} \mathcal{E} \subset \mathcal{K} \\ A_{F_0} \mathcal{E} \subset \mathcal{K} \\ \vdots \\ A_{F_0}^{k-1} \mathcal{E} \subset \mathcal{K} \end{cases} \end{aligned}$$

which is equivalent to (14) with $F = F_0$. \square

The following theorem can be proved in a similar quite easy way.

Theorem 2: There exists a static state feedback law $u(t) = Fx(t) + Gd(t)$, which solves the k th order Partial Modified Disturbance Rejection Problem if and only if

$$\mathcal{E} \subset \mathcal{V}^k + \mathcal{B} \quad (18)$$

where \mathcal{E} denotes $\text{Im} E$, \mathcal{B} denotes $\text{Im} B$ and \mathcal{V}^k denotes the subspace obtained at the k th step of (ISA) for (A, B, C) .

We shall now give some structural solutions in terms of structures at infinity.

IV. STRUCTURAL SOLVABILITY CONDITIONS

The following equivalent solvability condition will be expressed in terms of some equalities between integers, namely some orders of the zeros at infinity of (A, B, C) and $(A, [B \ E], C)$. Its main advantage comes from the fact that this structural information can be obtained through various ways, for instance from the Smith-McMillan Form at infinity, and thus is less dependent on a particular approach than Theorem 2.

Structural Solvability Condition for PMDRP(k)

Theorem 3: The k th order Partial Modified Disturbance Rejection Problem is solvable if and only if the orders of the zeros at infinity of both systems (A, B, C) and $(A, [B \ E], C)$, which are smaller than or equal to k , are the same.

To prove this theorem, we shall use the i th step of (ISA) for the combined system $(A, [B \ E], C)$, i.e.,

$$\begin{cases} \mathcal{V}_c^0 := \mathcal{X} \\ \mathcal{V}_c^i := \mathcal{K} \cap A^{-1}(\mathcal{B} + \mathcal{E} + \mathcal{V}_c^{i-1}), \quad i \geq 1 \end{cases}$$

which limit is \mathcal{V}_c^* , the supremal $(A, [B \ E])$ -invariant subspace contained in \mathcal{K} .

First note that (18) is obviously equivalent to

$$\mathcal{B} + \mathcal{E} + \mathcal{V}^i = \mathcal{B} + \mathcal{V}^i, \quad \forall i \in \{0, 1, \dots, k\}. \quad (19)$$

Let us introduce the integers

$$p_{ci} := \dim \left(\frac{(\mathcal{B} + \mathcal{E}) \cap \mathcal{V}_c^{i-1}}{(\mathcal{B} + \mathcal{E}) \cap \mathcal{V}_c^*} \right), \quad \forall i \geq 1$$

which characterize the orders of the zeros at infinity of $(A, [B \ E], C)$ (recall (7)), and remember that the list $\{p_{ci} - p_i\}$ characterizes the orders of the zeros at infinity of $(A, [B \ E], C)$ which are strictly smaller than i (remember Remark 1).

We shall prove that (19) is equivalent to

$$p_i - p_i = p_{ci} - p_{ci}, \quad \forall i \in \{1, 2, \dots, k+1\}. \quad (20)$$

Proof of Theorem 3: Note that if (19) is true, then

$$\mathcal{V}_c^i = \mathcal{V}^i, \quad \forall i \in \{0, 1, \dots, k+1\}. \quad (21)$$

Indeed, this is easily shown by induction:

Clearly, $\mathcal{V}_c^0 = \mathcal{V}^0$ and if $\mathcal{V}_c^{i-1} = \mathcal{V}^{i-1}$, then

$$\begin{aligned} \mathcal{V}_c^i &:= \mathcal{K} \cap A^{-1}(\mathcal{B} + \mathcal{E} + \mathcal{V}_c^{i-1}) \\ &= \mathcal{K} \cap A^{-1}(\mathcal{B} + \mathcal{E} + \mathcal{V}^{i-1}), \text{ since (21) holds for } i-1 \\ &= \mathcal{K} \cap A^{-1}(\mathcal{B} + \mathcal{V}^{i-1}), \text{ since (19) holds for } i-1 \\ &:= \mathcal{V}^i \end{aligned}$$

and thus $\mathcal{V}_c^i = \mathcal{V}^i$, $\forall i \in \{1, 2, \dots, k+1\}$.

From (9), (19) and (21), we have that

$$\begin{aligned} p_1 - p_i &= \dim \left(\frac{\mathcal{B}}{\mathcal{B} \cap \mathcal{V}^{i-1}} \right) = \dim \left(\frac{\mathcal{B} + \mathcal{V}^{i-1}}{\mathcal{V}^{i-1}} \right) \\ &= \dim \left(\frac{\mathcal{B} + \mathcal{E} + \mathcal{V}^{i-1}}{\mathcal{V}^{i-1}} \right) \\ &= \dim \left(\frac{\mathcal{B} + \mathcal{E} + \mathcal{V}_c^{i-1}}{\mathcal{V}_c^{i-1}} \right) \\ &= p_{ci} - p_{ci}, \quad \forall i \in \{1, 2, \dots, k+1\}. \end{aligned}$$

We shall now prove by induction the reverse part.

Indeed, (19) and (21) obviously hold for $i = 0$. Assume that (19) and (21) hold for $i - 1$, then

$$\begin{aligned} \mathcal{V}_c^i &:= \mathcal{K} \cap A^{-1}(\mathcal{B} + \mathcal{E} + \mathcal{V}_c^{i-1}) \\ &= \mathcal{K} \cap A^{-1}(\mathcal{B} + \mathcal{E} + \mathcal{V}^{i-1}), \text{ since (21) holds for } i-1 \\ &= \mathcal{K} \cap A^{-1}(\mathcal{B} + \mathcal{V}^{i-1}), \text{ since (19) holds for } i-1 \\ &:= \mathcal{V}^i \end{aligned}$$

and thus $\mathcal{V}_c^i = \mathcal{V}^i$, $\forall i \in \{0, 1, \dots, k+1\}$.

On the other hand, from (20) we have that

$$p_1 - p_{i+1} = p_{ci} - p_{ci+1}, \quad \forall i \in \{0, 1, \dots, k\}$$

then

$$\dim \left(\frac{\mathcal{B} + \mathcal{V}^i}{\mathcal{V}^i} \right) = \dim \left(\frac{\mathcal{B} + \mathcal{E} + \mathcal{V}_c^i}{\mathcal{V}_c^i} \right), \quad \forall i \in \{0, 1, \dots, k\}.$$

Since we have shown that $\mathcal{V}_c^i = \mathcal{V}^i$, $\forall i \in \{0, 1, \dots, k+1\}$, it follows that

$$\dim(\mathcal{B} + \mathcal{V}^i) = \dim(\mathcal{B} + \mathcal{E} + \mathcal{V}^i), \quad \forall i \in \{0, 1, \dots, k\}$$

which is equivalent to (19) and ends the proof. \square

The following corollary illustrates the obvious fact that exact rejection is just a particular case of partial rejection and brings back to some familiar structural results (see for instance [5] for the equivalent case (see Theorem 4.2) of model matching).

Corollary 1: The Exact Modified Disturbance Rejection Problem is solvable if and only if the k th order Partial Modified Disturbance Rejection Problem is solvable for the order $k = n_1$, where n_1 denotes the supremal order of the zeros at infinity of (A, B, C) .

Proof of Corollary 1: Obvious from Theorem 3, since Exact Rejection means Partial Rejection for any k and since the integers p_i are equal to zero for $i > n_1$. \square

We shall now deduce the corresponding structural result for PDRP(k).

For that purpose, consider the following "extended" perturbed system

$$\tilde{A} = \begin{bmatrix} A & I_n \\ 0 & 0 \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} 0 \\ B \end{bmatrix}, \quad \tilde{C} = [C \ 0], \quad \tilde{E} = \begin{bmatrix} E \\ 0 \end{bmatrix}. \quad (22)$$

The following lemma establishes some direct links between PMDRP(k) and PDRP(k).

Lemma 2 The k th order Partial Disturbance Rejection Problem for the system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + Ed(t) & t \geq 0 \\ y(t) = Cx(t) & t \geq 0 \end{cases}$$

is equivalent to the k th order Partial Modified Disturbance Rejection Problem for the extended system

$$\begin{cases} \dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) + \tilde{B}u(t) + \tilde{F}d(t) & t \geq 0 \\ \tilde{y}(t) = C\tilde{x}(t) & t \geq 0 \end{cases}$$

Proof of Lemma 2 Denote $\mathcal{V} = \mathcal{V}^k$ and let \mathcal{V} denote the step of (ISA) for (A, B, C) . The result is then an obvious consequence of the following relation

$$(\mathcal{V} + \mathcal{B}) \cap \mathcal{V} = \mathcal{V}$$

which proof is quite direct \square

Remark now that

$$C(sI - A)^{-1}B = s^{-1}C(sI - A)^{-1}B \quad (23)$$

and

$$C(sI - A)^{-1}F = C(sI - A)^{-1}F \quad (24)$$

This directly leads to the following

Structural Solvability Condition for PDRP(k)

Theorem 4 The k th order Partial Disturbance Rejection Problem is solvable if and only if the orders of the zeros at infinity of both systems (A, B, C) and $(A, [B, E], C)$ which are smaller than or equal to k are the same or equivalently if both transfer function matrices

$$T(s) = s^{-1}C(sI - A)^{-1}B$$

and

$$T(s) = [s^{-1}C(sI - A)^{-1}B \quad C(sI - A)^{-1}F]$$

have the same smallest orders of their zeros at infinity up to value k

To obtain the structural solution of the Partial Model Matching Problem we shall simply use the equivalence existing between this problem and the Partial Disturbance Rejection Problem

From Disturbance Rejection to Model Matching

The natural equivalence existing between Disturbance Rejection and Model Matching was shown in [3]. We shall here propose some alternative inspired from the nice equivalence existing between state feedback laws using dynamic extensions and general proper dynamic precompensators (see [4]) and from the fact that the use of a dynamic extension offers no extra possibility for Partial Disturbance Rejection. Let us define the following

Dynamic k th Order Partial Disturbance Rejection Problem

Do there exist some integer n_d and some dynamic extension (see [11])

$$x_d(t) = B_d u_d(t)$$

with $x_d \in \mathcal{V}_d$, $u_d \in \mathcal{U}_d$, $\dim(\mathcal{V}_d) = \dim(\mathcal{U}_d) = n_d$ such that the k th order Partial Rejection has a solution $u(t) = Fx_d(t)$ for the extended system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bx_d(t) + Fd(t) & t \geq 0 \\ y(t) = Cx(t) & t \geq 0 \end{cases} \quad (25)$$

where

$$x_d(t) = \begin{bmatrix} x_d(t) \\ x_d(t) \end{bmatrix} \in \mathcal{V}_d = \mathcal{V}_d \quad \mathcal{V}_d$$

$$u_d(t) = \begin{bmatrix} u_d(t) \\ u_d(t) \end{bmatrix} \in \mathcal{U}_d = \mathcal{U}_d \quad \mathcal{U}_d$$

$$A = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} \quad B = \begin{bmatrix} B & 0 \\ 0 & B \end{bmatrix} \quad C = [C \quad 0] \text{ and } I = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

Proposition 1 The k th order Dynamic Partial Disturbance Rejection Problem is solvable if and only if the (static) k th order Partial Disturbance Rejection Problem is solvable

Proof of Proposition 1 The proof is quite direct when using geometric arguments since it rests upon the following relationships

$$\mathcal{V}_d = \mathcal{V} \quad \mathcal{V}_d$$

where \mathcal{V} is the k th step of (ISA) for (A, B, C)

Then since $\mathcal{E} = \text{Im} F \subset \mathcal{V}$ it is quite obvious that

$$\mathcal{E} \subset \mathcal{V}^k \iff \mathcal{E} \subset \mathcal{V}_d^k$$

which ends the proof \square

This equivalence obviously remains true if the disturbance is available for the control law i.e. for solutions of the type

$$u(t) = Ix_d(t) + Gd(t) \quad (26)$$

Now it has been shown in [4] that the class of compensations like (26) applied to (25) is fully equivalent to the class of proper dynamic precompensators like

$$u(s) = C(s)d(s)$$

applied to the initial system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) & t > 0 \\ y(t) = Cx(t) & t \geq 0 \end{cases} \quad (27)$$

This equivalence immediately shows the following corollary

Corollary The k th order Modified Partial Disturbance Rejection Problem is solvable if and only if there exists a proper solution say $C(s)$ to the following equation

$$T(s)C(s) + T_d(s) = s^{-(k+1)}P(s) \quad (28)$$

where $P(s)$ is a proper rational transfer matrix and

$$T(s) = C(sI - A)^{-1}B \quad (29)$$

and

$$T_d(s) = C(sI - A)^{-1}F \quad (30)$$

Note that (28) is a particular case of the k th order Partial Model Matching Problem where $T(s)$ stands for the plant and $-I_d(s)$ for the model

We can then rephrase Corollary 2 as

Proposition 2 The k th order Modified Partial Disturbance Rejection Problem for (27) is equivalent to the k th order Partial Model Matching Problem with (29) as the plant and (30) as the model

As an immediate consequence of Lemma 2 we also have

Corollary 3 The k th order Partial Disturbance Rejection Problem is solvable if and only if there exists a strictly proper solution to (28)

Proof of Corollary 3: Use Lemma 2, i.e., just replace $T(s)$ by $s^{-1}T(s)$. This is equivalent to replacing $C(s)$ by $s^{-1}C(s)$, hence the result. \square

To show that the equivalence between Partial Disturbance Rejection and Partial Model Matching is complete, we shall quickly describe the reverse path.

Let us again consider the plant

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), & t \geq 0 \\ y(t) = Cx(t), & t \geq 0 \end{cases}$$

and the model

$$\begin{cases} \dot{x}_m(t) = A_m x_m(t) + B_m u_m(t), & t \geq 0 \\ y_m(t) = C_m x_m(t), & t \geq 0 \end{cases}$$

with their associated transfer matrices

$$T(s) := C(sI_n - A)^{-1}B$$

and

$$T_m(s) := C_m(sI_{n_m} - A_m)^{-1}B_m$$

respectively.

Let us introduce the following combination of the plant and of the model, and let us call it the combined system

$$\begin{cases} \dot{x}_c(t) = A_c x_c(t) + B_c u + E_c u_m(t), & t \geq 0 \\ y_c(t) = y(t) - y_m(t) = C_c x_c(t), & t \geq 0 \end{cases} \quad (31)$$

with

$$x_c(t) = \begin{bmatrix} x(t) \\ x_m(t) \end{bmatrix} \in \mathcal{X}_c \mathcal{X}_m$$

and

$$A_c = \begin{bmatrix} A & 0 \\ 0 & A_m \end{bmatrix}, B_c = \begin{bmatrix} B \\ 0 \end{bmatrix}, C_c = [C \quad -C_m] \text{ and } E_c = \begin{bmatrix} 0 \\ B_m \end{bmatrix}.$$

It directly follows from the previous discussion that looking for a proper dynamic compensator $u(s) = C(s)u_m(s)$, that solves the k th order Partial Model Matching Problem is equivalent to considering $u_m(s)$ as a (measured) disturbance in (31) and solving the corresponding k th order Modified Partial Disturbance Rejection Problem.

This complete equivalence allows us to directly deduce the following.

Geometric and Structural Solvability Conditions for PMMP(k)

Theorem 5. The k th order Partial Model Matching Problem has a proper solution $C(s)$ if and only if

$$\text{Im} \begin{bmatrix} 0 \\ B_m \end{bmatrix} \subset \text{Im} \begin{bmatrix} B \\ 0 \end{bmatrix} + \mathcal{V}_c^k$$

where \mathcal{V}_c^k is the k th step of (ISA) for the combined system.

An equivalent structural condition is the following.

The Partial Model Matching Problem has a proper solution $C(s)$

if and only if both transfer matrices $T(s)$ and $[T(s) : -T_m(s)]$ have the same orders of their zeros at infinity which are smaller than or equal to k .

V. CONCLUSION

We have considered here the solution to the Partial Model Matching Problem or, equivalently, the Partial Disturbance Rejection Problem. Lemma 2, Proposition 1, and Proposition 2 precisely describe the basic equivalences between these problems. Our geometric solution (Theorems 2 and 5) uses "classical" geometric tools, which are less involved than Generalized Dynamic Covers, as in [2]. The equivalent structural solution (Theorems 3 and 5) is expressed in terms of the orders of the zeros at infinity which are smaller than or equal to the desired order of the partial problem.

As argued in [2] this partial version of the problems is a very good intermediate to the exact one (see Corollary 1). It has to be noted that the same idea has been used in the context of systems with delays [8].

Further study have been devoted to some algebraic complementary study (with polynomial tools) and to stability requirements (which are compulsory for possible applications) [9]:

- It has been shown that there exists a stable solution to the Partial Model Matching Problem if and only if partial model matching is possible. Unlike in the exact case, there is no further constraint limitation imposed by unstable zeros for the problem to be solvable with stability.
- As is well known (see for instance [7]), exact disturbance rejection with stability has a static state feedback solution if and only if there exists a dynamic state feedback solution. As concerns the partial case, the situation is quite different. Indeed, it can be shown that there exists a dynamic stable solution to the Partial Disturbance Rejection Problem if and only if Partial Disturbance Rejection is actually solvable. However the characterization of static solutions (if any) is still open.
- The combination of partial disturbance rejection and optimal disturbance attenuation (i.e., minimization in H_∞ -terms of the effect of the disturbance on the output of the system while maintaining stability of the compensated plant) has been successfully achieved on some illustrative examples.

REFERENCES

- [1] C. Commault and J. M. Dion, "Structure at infinity of linear multivariable systems. A geometric approach," presented at the 20th IEEE Conf Decis. Contr., San Diego, CA, 1981.
- [2] E. Emre and L. H. Silverman, "Partial model matching of linear systems," *IEEE Trans Automat Contr.*, vol. AC-25, pp. 280-281, 1980.
- [3] E. Emre and M. L. J. Hautus, "A polynomial characterization of (A, B) -invariant and reachability subspaces," *SIAM J. Contr. Optim.*, vol. 18, no. 4, pp. 420-436, 1980.
- [4] V. Kučera and M. Malabre, "On various dynamic compensations," *Kybernetika*, vol. 19, no. 5, pp. 439-442, 1983.
- [5] M. Malabre, *Structure à l'Infini des Triplets Invariants Application à la Poursuite Parfaite de Modèle* (Lecture Notes in Control and Information Sciences), vol. 44. Berlin: Springer-Verlag, 1982.
- [6] —, "A complement about almost controllability subspaces," *Syst Contr. Lett.* no. 3, pp. 119-122, 1983.
- [7] M. Malabre and J. C. Martínez García, "The modified disturbance rejection problem with stability: A structural approach" in *Proc Second European Contr Conf (ECC'93)*, Groningen, The Netherlands, 1993, pp. 1119-1124.
- [8] M. Malabre and R. Rabah, "Structure at infinity, model matching and disturbance rejection for linear systems with delays," *Kybernetika*, vol. 29, no. 5, pp. 485-498, 1993.
- [9] J. C. Martínez García, M. Malabre, and V. Kučera, "The partial model matching problem with stability: Algebraic and structural solutions," *Laboratoire d'Automatique de Nantes, Ecole Centrale de Nantes, Nantes, France, Rapport Interne No. 93.20*, 1993.
- [10] A. I. G. Vardulakis, *Linear Multivariable Control Algebraic Analysis and Synthesis Methods*. New York: Wiley, 1991.
- [11] M. W. Wonham, *Linear Multivariable Control A Geometric Approach* 3rd ed. New York: Springer-Verlag, 1985.

Absolute Stability Criteria for Multiple Slope-Restricted Monotonic Nonlinearities

Wassim M. Haddad and Vikram Kapila

Abstract—Absolute stability criteria such as the classical Popov criterion guarantee stability for a class of sector-bounded nonlinearities. Although the sector restriction bounds the admissible class of the nonlinearities, the local slope of the nonlinearity may be arbitrarily large. In this paper we derive absolute stability criteria for multiple slope-restricted time-invariant monotonic nonlinearities. Like the Popov criterion, in the single-input/single-output case our results provide a simple graphical interpretation involving a straight line in a modified Popov plane.

I. INTRODUCTION

Absolute stability theory guarantees stability of feedback systems whose forward path contains a dynamic linear time invariant system and whose feedback path contains a memoryless (possibly time varying) nonlinearity. These stability criteria are generally stated in terms of the linear system and apply to every element of a specified class of nonlinearities. Hence absolute stability theory provides sufficient conditions for robust stability with a given class of uncertain elements [5]–[17].

The literature on absolute stability is extensive. A convenient way to distinguish these results is to focus on the allowable class of feedback nonlinearities. Specifically the small gain, positivity, and circle theorems guarantee stability for arbitrarily time varying nonlinearities, whereas the Popov criterion does not. This is not surprising since the Lyapunov function upon which the small gain, positivity, and circle theorems are based is a fixed quadratic Lyapunov function which permits arbitrary time variation of the nonlinearity [5]. Alternatively the Popov criterion is based on a Lur'e–Postnikov Lyapunov function which explicitly depends on the nonlinearity thereby restricting its allowable time variation.

Further refinements of absolute stability criteria developed in [2], [12], [14], [19] restrict consideration to sector bounded time invariant nonlinearities that are monotonic or odd monotonic and are predicted on extended Lur'e–Postnikov Lyapunov functions [9], [14]. To further restrict the allowable class of feedback nonlinearities the authors in [4], [17], [22]–[24] develop absolute stability criteria by constraining the local slope of the nonlinearity. These classical absolute stability results extend the Popov criterion for sector bounded time invariant nonlinear functions to monotonic and odd monotonic nonlinearities by constructing stability multipliers that effectively place less restrictive conditions on the linear part of the system. However, as a result of the more involved multiplier construction the resulting frequency domain conditions do not provide a simple graphical test as in the case of the Popov criterion.

In recent research [10], [16], [18] a new absolute stability criterion for locally slope restricted nonlinearities involving a simple modification to the Popov multiplier was developed. Specifically it is shown in [18] that replacing the Popov multiplier $Z(s) = 1 + \lambda s$ by the new multiplier $1 + \lambda s^{-1}$ and requiring the frequency domain condition $\mu^{-1} + (1 + \lambda s^{-1})G(s)$ be positive real where

$G(s)$ represents the transfer function of the linear dynamic system and μ is a bound on the local slope of the feedback nonlinearity provides a sufficient condition for the absolute stability for systems with a monotonic time invariant nonlinear element in the feedback path. As noted in [10], [16], however, the statement as well as the proof of the results reported in [18] were far from convincing. It should further be noted that the authors in [10] consider extensions to several differentiable nonlinearities in the feedback loop using an involved method based on integral indices along with stability inequalities that arise from the frequency domain condition. To provide connections between the proposed absolute stability condition and robust controller analysis using the parameterized Lyapunov function framework developed in [5], in this paper we extend the results of [18] to multiple slope restricted monotonic nonlinearities as well as construct explicit Lyapunov functions along with providing the underlying Yakubovich–Kalman–Popov conditions needed to present a concise statement of these results. Specifically an extended notion of a kinetic Lyapunov function [3] is used to show asymptotic stability of the nonlinear feedback system given by

$$\dot{x}(t) = Ax(t) - B\phi(y) \quad y(t) = Cx(t)$$

where $\phi(\cdot)$ is a time invariant sector bounded memoryless nonlinearity. That is, instead of finding the condition for the state variables $x(t)$ to approach the zero equilibrium point, sufficient conditions for $x(t)$ and the output $y(t)$ to approach zero are found. Obviously if the system is observable and $x(t)$ and $y(t)$ approach zero, the system arrives at one of its equilibrium points, and the two results are equivalent if the equilibrium point is unique. Finally, in the single input/single output (SISO) case, we show that the resulting frequency domain condition has a simple graphical interpretation involving a straight line in a modified Popov plane.

II. MATHEMATICAL PRELIMINARIES

In this section we establish definitions, notation, and several key lemmas. Let \mathbb{R} and \mathbb{C} denote the real and complex numbers. Let $(\cdot)^T$ and $(\cdot)^*$ denote transpose and complex conjugate transpose. Let I_n or I denote the $n \times n$ identity matrix and let 0 denote the $n \times n$ zero matrix. Furthermore, $M \geq 0$ ($M > 0$) denotes the fact that the Hermitian matrix M is nonnegative (positive) definite. Let $n(s)$ and $d(s)$ be polynomials in s with real coefficients. A function $q(s)$ of the form $q(s) = n(s)/d(s)$ is called a rational function. The function $q(s)$ is called proper (respectively strictly proper) if $\deg n(s) \leq \deg d(s)$ (respectively $\deg n(s) < \deg d(s)$) where \deg denotes the degree of the polynomial. In this paper a real rational matrix function is a matrix whose elements are rational functions with real coefficients. Furthermore, a transfer function $G(s)$ is called proper (respectively strictly proper) if every element of $G(s)$ is proper (respectively strictly proper). Finally, an asymptotically stable transfer function is a transfer function each of whose poles is in the open left half plane. The space of asymptotically stable transfer functions is denoted by \mathcal{RH}_∞ , i.e., the real rational subset of \mathcal{H}_∞ . Let

$$G(s) \sim \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

denote a state space realization of a transfer function $G(s)$ that is $G(s) = C(sI - A)^{-1}B + D$. The notation $\overset{\sim}{\sim}$ is used to denote a minimal realization. In addition, the parahermitian conjugate $G^*(s)$

Manuscript received January 30, 1993; revised April 18, 1994. This research was supported in part by the National Science Foundation Research Grants ECS 9109558 and ECS 9350181.

The authors are with the School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0150, USA.
IEEE Log Number 9407219.

of $G(s)$ has the realization

$$G(s) \sim \begin{bmatrix} -A^T & C^T \\ -B^T & D^T \end{bmatrix}.$$

Furthermore, the Hermitian part of G is given by $\text{He } G = \frac{1}{2}(G + G^*)$.

A square transfer function $G(s)$ is called positive real [1, p. 216] if 1) all poles of $G(s)$ lie in the closed left-half plane, and 2) $\text{He } G(s)$ is nonnegative definite for $\text{Re } s > 0$. A square transfer function $G(s)$ is called strictly positive real [21] if 1) $G(s)$ is asymptotically stable and 2) $\text{He } G(j\omega)$ is positive definite for all real ω . Recall that a minimal realization of a positive real transfer function is stable in the sense of Lyapunov, while a minimal realization of a strictly positive real transfer function is asymptotically stable.

For notational convenience we will omit all matrix dimensions throughout the paper and assume that all quantities have compatible dimensions. Furthermore, in this paper, $G(s)$ will denote an $m \times m$ transfer function with input $u \in \mathbb{R}^m$, output $y \in \mathbb{R}^m$, and internal state $x \in \mathbb{R}^n$. Next, we state the strict positive real lemma used to characterize strict positive realness in the state-space setting.

Lemma 2.1 (Strict Positive Real Lemma [20]):

$$G(s) \stackrel{\text{min}}{\sim} \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

is strictly positive real if and only if there exist matrices P , L , and W with P positive definite such that

$$0 = A^T P + P A + L^T L, \quad (2.1)$$

$$0 = B^T P - C^T + W^T L, \quad (2.2)$$

$$0 = D + D^T - W^T W, \quad (2.3)$$

are satisfied, the pair (A, L) is observable, and $\text{rank } \hat{G}(j\omega) = m$, $\omega \in \mathbb{R}$, where

$$\hat{G}(s) \stackrel{\text{min}}{\sim} \begin{bmatrix} A & B \\ L & W \end{bmatrix}.$$

Finally, we state a key lemma involving controllability of an augmented pair.

Lemma 2.2 [10]: Given a triple (A, B, C) , if (A, B) is controllable and $\det A \neq 0$, then

$$\left(\begin{bmatrix} A & 0 \\ C & 0 \end{bmatrix}, \begin{bmatrix} B \\ 0 \end{bmatrix} \right)$$

is controllable if and only if $\det C A^{-1} B \neq 0$.

III. ABSOLUTE STABILITY CRITERION FOR MULTIPLE SLOPE-RESTRICTED NONLINEARITIES

In this section we consider the absolute stability problem for a class Φ of locally slope restricted monotonic time-invariant nonlinearities $\phi: \mathbb{R}^m \rightarrow \mathbb{R}^m$. Specifically, given

$$G(s) \stackrel{\text{min}}{\sim} \begin{bmatrix} A & B \\ C & 0 \end{bmatrix}$$

we derive conditions that guarantee global asymptotic stability of the negative feedback interconnection of $G(s)$ and ϕ for all $\phi \in \Phi$. Note that the negative feedback interconnection of $G(s)$ and $\phi(\cdot)$ has the state-space representation

$$\dot{x}(t) = A x(t) - B \phi(y), \quad (3.1)$$

$$y(t) = C x(t), \quad (3.2)$$

To state our main result, the following definitions are needed. Let $\mu \in \mathbb{R}^{m \times m}$ be a positive definite diagonal matrix. Next, define the set Φ of allowable nonlinearities ϕ by

$$\Phi \triangleq \{ \phi: \mathbb{R}^m \rightarrow \mathbb{R}^m : \phi(y) = [\phi_1(y_1), \dots, \phi_m(y_m)]^T, \\ \phi(\cdot) \text{ is differentiable,} \\ 0 < \phi'_i(y_i) < \mu_{ii}, \quad i = 1, \dots, m, y \in \mathbb{R}^m \}. \quad (3.3)$$

Note that the nonlinear functions considered, $\phi \in \Phi$, have decoupled components but unlike the multivariable extensions of the Popov criterion [5], [11] we assume a local slope constraint on the nonlinearities. In the scalar case, $m = 1$, ϕ satisfies the usual local slope condition $0 < \phi'(y) < \mu$, $y \in \mathbb{R}$.

For the statement of the main result define

$$A_a \triangleq \begin{bmatrix} A & 0_{n \times m} \\ C & 0_m \end{bmatrix}, \quad B_a \triangleq \begin{bmatrix} B \\ 0_m \end{bmatrix}, \\ C_a \triangleq [0_{m \times n} \ I_m], \quad S \triangleq [C^T \ 0_m].$$

Theorem 3.1. Let

$$G(s) \stackrel{\text{min}}{\sim} \begin{bmatrix} A & B \\ C & 0 \end{bmatrix}$$

be asymptotically stable, let $N \triangleq \text{diag } [N_1, N_2, \dots, N_m]$ be nonnegative-definite, assume $\det C A^{-1} B \neq 0$, and define

$$Z(s) \triangleq I + N s^{-1}. \quad (3.4)$$

Then

$$\mathcal{G}(s) \triangleq \mu^{-1} + Z(s)G(s) \quad (3.5)$$

is strictly positive real if and only if there exist matrices P , L , and W with P positive definite satisfying

$$0 = A_a^T P + P A_a + L^T L, \quad (3.6)$$

$$0 = B_a^T P - N C_a - S + W^T L, \quad (3.7)$$

$$0 = 2\mu^{-1} - W^T W. \quad (3.8)$$

In this case

$$V(\dot{x}, y) = \begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix}^T P \begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix} \\ + 2 \sum_{i=1}^m \int_0^{y_i} N_i \phi'_i(\sigma) \sigma d\sigma \quad (3.9)$$

where $y(t) = C x(t)$, is a Lyapunov function that guarantees that the negative feedback interconnection of $G(s)$ and $\phi(\cdot)$ is globally asymptotically stable for all $\phi \in \Phi$.

Proof: First, define $z(t) \triangleq \dot{x}(t)$ so that

$$\dot{z}(t) = A z(t) - B \phi'(y) \dot{y}, \quad (3.10)$$

$$\dot{y}(t) = C z(t) \quad (3.11)$$

and

$$\dot{x}_a(t) = A_a x_a(t) - B_a f(y) \quad (3.12)$$

where

$$x_a(t) \triangleq \begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix}, \quad f(y) \triangleq \phi'(y) \dot{y}(t).$$

Furthermore, note that in this case $y(t) = C_a x_a(t)$ and $\dot{y}(t) = S x_a(t)$. Now, since (A, B, C) is minimal, (A, B) is controllable. Hence, it follows from Lemma 2.2 that if $\det C A^{-1} B \neq 0$ then (A_a, B_a) is also controllable.

Next, we show that (3.6)–(3.8) imply that $\mathcal{G}(s)$ is strictly positive real. To do this, add and subtract $j\omega P$ to and from (3.6) to obtain

$$0 = (-j\omega I - A_n)^T P + P(j\omega I - A_n) - L^T L. \quad (3.13)$$

Now, forming $B_n^T (-j\omega I - A_n)^{-T}$ (3.13) $(j\omega I - A_n)^{-1} B_n$ and using (3.7) we obtain

$$\begin{aligned} & [NC_n + S - W^T L](j\omega I - A_n)^{-1} B_n \\ & + B_n^T (-j\omega I - A_n)^{-T} [NC_n + S - W^T L]^T \\ & = B_n^T (-j\omega I - A_n)^{-T} L^T L (j\omega I - A_n)^{-1} B_n. \end{aligned} \quad (3.14)$$

Adding and subtracting $W^T W$ to and from (3.14), using (3.8), and grouping terms yields

$$\begin{aligned} & [NC_n + S](j\omega I - A_n)^{-1} B_n \\ & + B_n^T (-j\omega I - A_n)^{-T} [NC_n + S]^T + 2\mu^{-1} \\ & = [W + L(j\omega I - A_n)^{-1} B_n]^* \\ & \cdot [W + L(j\omega I - A_n)^{-1} B_n]. \end{aligned} \quad (3.15)$$

Next, using the identities

$$\begin{aligned} (j\omega I - A_n)^{-1} &= \begin{bmatrix} (j\omega I - A) & 0 \\ -C & j\omega I \end{bmatrix}^{-1} \\ &= \begin{bmatrix} (j\omega I - A)^{-1} & 0 \\ (j\omega)^{-1} C(j\omega I - A)^{-1} & (j\omega)^{-1} I \end{bmatrix} \end{aligned}$$

$$S(j\omega I - A_n)^{-1} B_n = G(j\omega),$$

$$C_n(j\omega I - A_n)^{-1} B_n = \frac{G(j\omega)}{j\omega}.$$

it follows from (3.15) and the rank condition in ii) of Lemma 2.1 that $\text{Re } \mathcal{G}(j\omega) > 0$. Hence $\mathcal{G}(s)$ is strictly positive real.

Conversely, assuming that $\mathcal{G}(s)$ is strictly positive real, spectral factorization theory guarantees the existence of a spectral factor $\Lambda(s)$ such that $\mathcal{G}(s) + \mathcal{G}^*(s) = \Lambda^*(s)\Lambda(s)$, where $\Lambda^{\pm 1}(s) \in RH_\infty$. The existence of P , L , and W with P positive definite satisfying (3.6)–(3.8) now follows from standard algebraic state-space realization manipulations.

Alternatively, the result follows from a direct consequence of Lemma 2.1 by noting that $\mathcal{G}(s)$ has a minimal realization given by

$$\mathcal{G}(s) \stackrel{\text{min}}{\sim} \left[\begin{array}{c|c} A_n & B_n \\ \hline NC_n + S & \mu^{-1} \end{array} \right].$$

Next, for $\phi \in \Phi$ consider the Lyapunov function candidate (3.9). First note that using integration by parts the integral term in (3.9) is equivalent to

$$\sum_{i=1}^m \int_0^{y_i} N_i \phi_i(\sigma) \sigma d\sigma = \sum_{i=1}^m \left\{ N_i \phi_i(y_i) y_i - \int_0^{y_i} N_i \phi_i(\sigma) d\sigma \right\}.$$

Now, since $\phi'(y) > 0$, for all $\phi \in \Phi$, it follows that $\sum_{i=1}^m \{N_i \phi_i(y_i) y_i - \int_0^{y_i} N_i \phi_i(\sigma) d\sigma\} > 0$. Furthermore, since V is positive definite, it follows that $\dot{V}(x_n)$ is positive definite. The corresponding Lyapunov derivative is given by

$$\begin{aligned} \dot{V}(x_n) &= x_n^T [A_n^T P + P A_n] x_n \\ &\quad - f^T(y) [B_n^T P - N C_n] x_n \\ &\quad - x_n^T [B_n^T P - N C_n]^T f(y). \end{aligned} \quad (3.16)$$

Next, it follows from (3.3) that $\phi'(y)(I - \mu^{-1} \phi'(y)) > 0$, for all $\phi \in \Phi$. Hence, $2 \dot{y}^T \phi'(y)(I - \mu^{-1} \phi'(y)) \dot{y} \geq 0$. Now, adding

and subtracting $2 \dot{y}^T \phi'(y)(I - \mu^{-1} \phi'(y)) \dot{y}$ to and from (3.16) and grouping terms yields

$$\begin{aligned} \dot{V}(x_n) &= x_n^T [A_n^T P + P A_n] x_n \\ &\quad - f^T(y) [B_n^T P - N C_n - S] x_n \\ &\quad - x_n^T [B_n^T P - N C_n - S]^T f(y) \\ &\quad - 2 f^T(y) \mu^{-1} f(y) - 2 \dot{y}^T \phi'(y) \\ &\quad \cdot (I - \mu^{-1} \phi'(y)) \dot{y} \end{aligned} \quad (3.17)$$

or, using (3.6)–(3.8), yields

$$\begin{aligned} \dot{V}(x_n) &= -[L x_n - W f(y)]^T [L x_n - W f(y)] \\ &\quad - 2 \dot{y}^T \phi'(y) (I - \mu^{-1} \phi'(y)) \dot{y}. \end{aligned} \quad (3.18)$$

Since $2 \dot{y}^T \phi'(y) (I - \mu^{-1} \phi'(y)) \dot{y} \geq 0$, it follows that $\dot{V}(x_n) \leq 0$, which proves stability in the sense of Lyapunov.

To show global asymptotic stability we need to show that $\dot{V}(x_n) = 0$ implies that $x = 0$. Note that $\dot{V}(x_n) = 0$ implies that $\dot{y}(t) = 0$, $t \geq 0$, and hence $f(y) = 0$ and $L x_n(t) = 0$. Furthermore, in this case $\dot{x}_n(t) = A_n x_n(t) - B_n f(y) = A_n x_n(t)$. Thus, using $\dot{x}_n(t) = A_n x_n(t)$, $L x_n(t) = 0$, and the observability of (A_n, L) , it follows from the PBH test that $x_n(t) = 0$, $t \geq 0$, which further implies, since (A, C) is observable, that $x(t) = 0$, $t \geq 0$. Thus, the only solution satisfying $\dot{V}(x, y) = 0$ is the $x(t) = 0$, $t \geq 0$, solution and hence it follows from the LaSalle's theorem [20] that global asymptotic stability holds. \square

Theorem 3.1 presents a generalization of Theorem 1 of [18] to the case of multivariable plants containing an arbitrary number of memoryless time-invariant slope-restricted monotonic nonlinearities.

The form of $\mathcal{G}(s)$ given by (3.5) is standard in the classical absolute stability theory [14] in which $Z(s)$ is a stability multiplier that distinguishes the class of the allowable feedback nonlinearities. As mentioned in the Introduction specific cases include memoryless time-invariant nonlinearities [15], monotonic and odd monotonic nonlinearities [9], [12]–[14], and locally slope restricted nonlinearities [4], [22]–[24]. A key difference between the results in Theorem 3.1 and the classical theory on monotonic and odd monotonic nonlinearities [9], [12]–[14] is that the multiplier $Z(s)$ in (3.4) involves a simple twist on the Popov multiplier in contrast to the more involved positive real multipliers involving partial fraction expansions of driving point impedances of resistor-inductor (\mathcal{RL}) and resistor-capacitor (\mathcal{RC}) combinations which exhibit interlacing pole-zero patterns on the negative real axis [9], [12], [13] and non-positive real plant-dependent multipliers [4], [7], [23].

In the SISO case, the frequency domain condition in Theorem 3.1 has an interesting geometric interpretation. Specifically, setting $G(j\omega) = x + jy$, $\text{Re } \mathcal{G}(j\omega) > 0$ is equivalent to

$$\frac{1}{\mu} + x + \frac{N}{\omega} y > 0. \quad (3.19)$$

Condition (3.19) is a frequency domain stability criterion with a graphical interpretation in a modified Popov plane, involving $\text{Re } G$ and $\omega^{-1} \text{Im } G$, in terms of a straight line with a real axis-intercept $-1/\mu$ and slope $-1/N$.

Since the condition presented in Theorem 3.1 is only sufficient for absolute stability a natural question that arises is for what class of systems will this criterion give less conservative predictions over the classical Popov criterion. In order to address this question first recall that the effect that the Popov multiplier $1 + Ns$ has on the Nyquist plot is to rotate each point on the Nyquist plot in the counter clockwise direction. Hence, if the Nyquist plot of the plant transfer function $G(s)$ enters the second quadrant then it is clear that there does not exist an N such that $(1 + Ns)G(s)$ is positive real. Thus, in

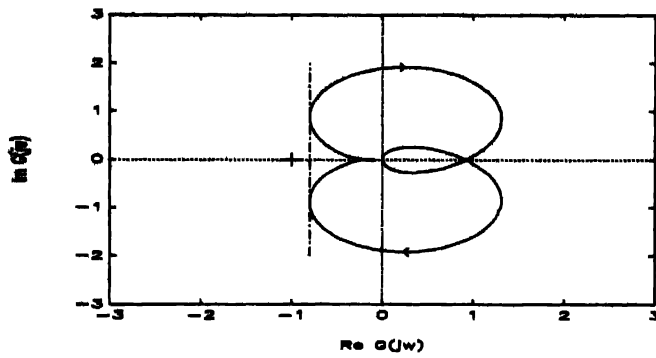


Fig. 1. Positive real analysis

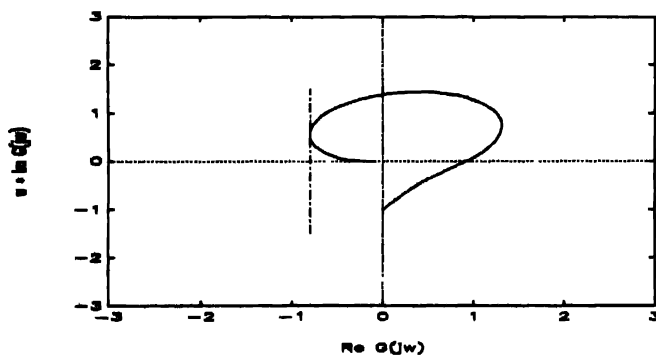


Fig. 2. Popov analysis.

this case, the Popov criterion does not provide any improvement over the positive real test (see the example in Section IV). Alternatively, since the effect of the proposed multiplier $1 + Ns^{-1}$ is to rotate each point on the Nyquist plot in the clockwise direction, the criterion in Theorem 3.1 will always give less conservative predictions over the Popov criterion when the Nyquist plot of $G(s)$ resides in the first and second quadrants. For example, since the Nyquist plot of the class of third-order transfer functions given by

$$G(s) = \frac{s^2}{a_0 s^3 + a_1 s^2 + a_2 s + a_3}$$

where $a_0 > 0$, $a_1 > 0$, $a_2 > 0$, $a_3 > 0$, and $a_2 a_1 > a_0 a_3$ will always enter the second quadrant, the proposed criterion would give less conservative predictions over the Popov criterion. Of course, using similar arguments as above, if the Nyquist plot of $G(s)$ resides in the third and fourth quadrants then the proposed criterion would not give any improvement over the positivity criterion while the Popov criterion would give less conservative predictions. Hence, the utility of the proposed criterion is when the Popov criterion fails. Finally, it should be noted that the more general class of multipliers consisting of the \mathcal{RL} and the \mathcal{RC} class [9], [12], [13] place less restrictive conditions on $G(s)$ and hence allow the Nyquist plot to reside in all four quadrants. As a result of the more involved multiplier construction, however, the resulting frequency domain conditions provide a complex graphical interpretation involving frequency-dependent off-axis circles in the Nyquist plane [9]. Alternatively, if the classical off-axis circle criterion is used where a single bounding circle in the Nyquist plane is employed [14] as opposed to a family of frequency dependent circles then conservatism will be introduced in the stability predictions.

Remark 3.1 Note that the class Φ of nonlinearities becomes larger as μ increases. In fact, as μ increases the strict positive real condition (3.5) becomes more difficult to satisfy, as expected. Furthermore,

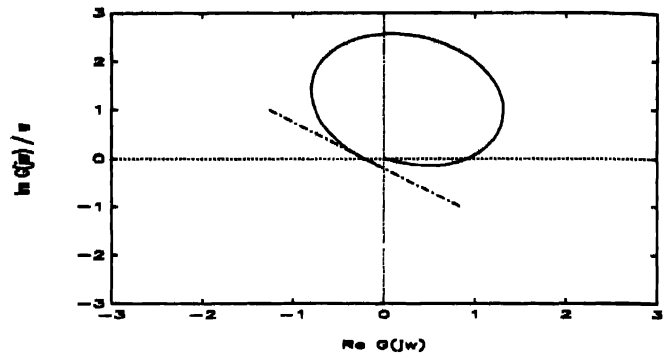


Fig. 3. Modified Popov analysis

even though the frequency domain condition in Theorem 3.1 does not involve an explicit sector constraint on the nonlinearities $\phi(y)$, the requirement that $\phi \in \Phi$ implies that $0 < \phi_i(y_i)y_i < \mu y_i^2$, $y_i \in \mathbb{R}$, $i = 1, \dots, m$.

Next, we partially relax the assumption $\det CA^{-1}B \neq 0$ and provide an alternative Lyapunov function construction for the absolute stability criterion given in Theorem 3.1. The following result does not require a system augmentation of the form (3.12), however, in this case we assume that every element of the stable transfer function $G(s)$ has at least one zero at the origin, i.e., $G(0) = 0$.

Theorem 3.2 Let

$$G(s) \approx \begin{bmatrix} A & B \\ C & 0 \end{bmatrix}$$

be asymptotically stable, let $N \triangleq \text{diag}[N_1, N_2, \dots, N_m]$ be nonnegative-definite, assume that $G(0) = 0$, and define

$$Z(s) \triangleq I + Ns^{-1}. \quad (3.20)$$

Then

$$\tilde{G}(s) \triangleq \mu^{-1} + Z(s)G(s) \quad (3.21)$$

is strictly positive real, if and only if there exists matrices P , L , and W with P positive definite satisfying

$$0 = A^T P + P A + L^T L, \quad (3.22)$$

$$0 = B^T P - N C A^{-1} - C^T + W^T L, \quad (3.23)$$

$$0 = 2\mu^{-1} - W^T W. \quad (3.24)$$

In this case

$$V(\dot{x}) = \dot{x}^T P \dot{x} + 2 \sum_{i=1}^m \int_0^{y_i} N_i \phi_i(\sigma) \sigma d\sigma \quad (3.25)$$

where $y(t) = Cx(t)$, is a Lyapunov function that guarantees that the negative feedback interconnection of $G(s)$ and ϕ is globally asymptotically stable for all $\phi \in \Phi$.

Proof The proof is similar to the proof of Theorem 3.1. \square

Remark 3.2 Note that since $G(s)$ is asymptotically stable A^{-1} exists. Furthermore, in order to construct the proof of Theorem 3.2, it is helpful to note that since every element of the stable transfer function $G(s)$ has at least one zero at the origin, $CA^{-1}B = 0$ and $CA^{-1}(sI - A)^{-1}B = \frac{G(s)}{s}$.

IV. ILLUSTRATIVE NUMERICAL EXAMPLE

For illustrative purposes we consider the stable linear system with transfer function

$$G(s) = \frac{s^2 - 0.2s - 0.1}{s^3 + 2s^2 + s + 1}. \quad (4.1)$$

the closed-loop system (3.1) with linear uncertainty $\phi(y) = Fy$ is asymptotically stable for $0 \leq F \leq 4.6$. The Nyquist plot for the linear system is shown in Fig. 1. Hence, it follows from the positivity theorem that the linear system is asymptotically stable for all time invariant monotone nonlinearities in the sector $[0, 1/24]$. Fig. 2 shows the corresponding Popov plot which, in this case, gives a Popov sector $[0, 1/24]$. Hence, since the Nyquist plot of $G(s)$ enters the second quadrant the Popov criterion does not provide any improvement over the positive real test. Finally, using Theorem 3.1 we construct a modified Popov plot shown in Fig. 3. Using the modified Popov exclusionary half plane graphical test given by (3.19) the absolute stability sector is now found to be $[0, 4.6]$ which is a significant improvement over the positivity and Popov sectors for time invariant monotone feedback nonlinearities.

CONCLUSION

In this paper we extended the SISO absolute stability criterion for locally slope restricted monotonic nonlinearities developed in [18] to multivariable systems containing an arbitrary number of monotonic slope bounded nonlinearities. Specifically, explicit Lyapunov functions along with extended Yakubovich-Kalman-Popov conditions are given. These results can be used to synthesize robust feedback controllers in the spirit of [6], [8], [9].

REFERENCES

- [1] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis: A Modern Systems Theory Approach*, Englewood Cliffs, NJ: Prentice Hall, 1973.
- [2] R. W. Brockett and J. I. Willms, "Frequency domain stability criteria I and II," *IEEE Trans Automat Contr*, vol. AC-10, pp. 255-261, 1965.
- [3] S. S. I. Chung, "Kinetic Lyapunov function for stability analysis of nonlinear control systems," *J Basic Engineering Trans ASME*, vol. 83, D, pp. 91-94, 1961.
- [4] A. G. Dewey and I. I. Jury, "A stability inequality for a class of nonlinear feedback systems," *IEEE Trans Automat Contr*, vol. AC-11, pp. 54-62, 1966.
- [5] W. M. Haddad and D. S. Bernstein, "Explicit construction of quadratic Lyapunov functions for the small gain positivity circle and Popov Theorems and their application to robust stability. Part I: Continuous time theory," *Int J Robust and Nonlinear Control*, vol. 3, pp. 313-339, 1993.
- [6] —, "Parameter dependent Lyapunov functions, constant real parameter uncertainty and the Popov criterion in robust analysis and synthesis," in *Proc IEEE Conf Dec Contr*, Brighton, UK, 1991, pp. 2274-2279, 2632-2633.
- [7] —, "Off axis absolute stability criteria and μ bounds involving non positive real plant dependent multipliers for robust stability and performance with locally slope restricted monotonic nonlinearities," in *Proc Amer Contr Conf*, San Francisco, CA, 1993, pp. 2790-2794.
- [8] —, "Parameter dependent Lyapunov functions and the Popov criterion in robust analysis and synthesis," *IEEE Trans Automat Contr*, to appear.
- [9] W. M. Haddad, J. P. How, S. R. Hall, and D. S. Bernstein, "Extensions of mixed μ bounds to monotonic and odd monotonic nonlinearities using absolute stability theory," in *Proc IEEE Conf Decis Contr*, Tucson, AZ, 1992, pp. 2813-2823.
- [10] A. Halanay and V. Rasvan, "Absolute stability of feedback systems with several differentiable nonlinearities," *Int J Systems Sci*, vol. 22, pp. 1911-1927, 1991.
- [11] J. B. Moore and B. D. O. Anderson, "A generalization of the Popov criterion," *J Franklin Inst*, vol. 285, pp. 488-492, 1968.
- [12] K. S. Narendra and C. P. Neuman, "Stability of a class of differential equations with a single monotonic nonlinearity," *SIAM J Control Optim*, vol. 4, pp. 295-308, 1966.

- [13] —, "Stability of continuous time dynamical systems with nonlinearities," *AIAA Journal*, vol. 5, pp. 2021-2027, 1967.
- [14] K. S. Narendra and J. H. Taylor, *Frequency Domain Criteria for Absolute Stability*, New York: Academic Press, 1973.
- [15] V. M. Popov, "Absolute stability of nonlinear systems," *IEEE Trans Automat and Remote Contr*, vol. 22, pp. 851-875, 1966.
- [16] V. Rasvan, "New results and applications of the frequency domain criteria to absolute stability of nonlinear systems," *Quarterly J of Differential Equations*, vol. 53, pp. 577-594, 1988.
- [17] M. G. Salomonov, "Stability of interconnected systems having slope bounded nonlinearities," in *Proc Sixth Int Conf Anal Optim*, Nice, France, 1984, pp. 275-287.
- [18] V. Singh, "A stability inequality for nonlinear feedback systems with slope restricted nonlinearity," *IEEE Trans Automat Contr*, vol. AC-29, pp. 743-744, 1984.
- [19] M. A. I. Thathachar and M. D. Srinath, "Some aspects of the Lur'e problem," *IEEE Trans Automat Contr*, vol. AC-12, pp. 451-453, 1967.
- [20] M. Vidyasagar, *Nonlinear Systems Analysis*, Englewood Cliffs, NJ: Prentice Hall, 1993.
- [21] J. I. Wen, "Time domain and frequency domain conditions for strict positive realness," *IEEE Trans Automat Contr*, vol. AC-33, pp. 985-992, 1988.
- [22] V. A. Yakubovich, "The method of matrix inequalities in the theory of the stability of nonlinear control systems II: Absolute stability in a class of nonlinearities with a condition on the derivative," *Automat Remote Contr*, vol. 26, pp. 577-592, 1965.
- [23] —, "Frequency conditions for the absolute stability and dissipativity of control systems with a single differentiable nonlinearity," *Sov Math*, vol. 6, pp. 98-101, 1965.
- [24] G. Zames and P. L. Falb, "Stability conditions for systems with monotone and slope restricted nonlinearities," *SIAM J Control Optim*, vol. 4, pp. 89-108, 1968.

Simultaneous Disturbance Rejection and Regular Row by Row Decoupling With Stability: A Geometric Approach

Juan Carlos Martinez Garcia and Michel Malabre

Abstract—The simultaneous disturbance rejection problem and regular row by row decoupling problem with stability is solved here through a geometric approach. It is shown in this paper that the combined problem has a solution if and only if each problem, separately, has a solution.

1 INTRODUCTION

As far as its geometric setting is concerned, the combined problem of disturbance rejection and input-output decoupling for linear time invariant systems by static state feedback has been first discussed in [1] and [3] almost 20 years ago. This problem has been recently revisited in [9] while the structural approach has been used in [2] in order to obtain easy to verify solvability conditions.

As the main result of the research referenced above, a nice property has been found. Indeed, when no stability constraint is imposed it

Manuscript received December 21, 1993; revised May 6, 1994. This work was supported in part by National Council of Science and Technology of Mexico, the Advanced Studies and Research Center of the IPN of Mexico and ESPRIT Basic Research Project No. 8924 (SESDIP).

J. C. Martinez Garcia was with the LAN, URA CNRS 823, Ecole Centrale de Nantes, Université de Nantes, 1 rue de la Noë, F-44072 Nantes Cedex 03, France, and is now with CINESTAV, Mexico.

M. Malabre is with the LAN, URA CNRS 823, Ecole Centrale de Nantes, Université de Nantes, 1 rue de la Noë, F-44072 Nantes Cedex 03, France. IEEE Log Number 9407221.

has been established that the combined problem is solvable if and only if each problem, separately, has a solution. Intuitively, when the internal stability of the closed-loop system is required, this separation solvability should still hold. To our knowledge, this result has never been established. The aim of this paper is to bridge that gap through a geometric approach.

First of all, we shall present some basic tools and the geometric solvability conditions of both the regular row by row decoupling problem with stability and the disturbance rejection problem with stability, where the notion of supremal internally stabilizable (A, B) -invariant subspace will play an important role.

In the second part of the paper we shall present the solution of the combined problem, proving that regular row by row decoupling with stability can be obtained using a static state feedback control law chosen in a particular family (built up from a particular set of internally stabilizable (A, B) -invariant subspaces), which also rejects automatically the disturbance, if this latter problem is solvable. This "separation property" is proved under the assumption that the system without disturbance is controllable. This is indeed the usual assumption when studying decoupling.

II. BASIC CONCEPTS

Let us consider a linear time-invariant system described by

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), & t \geq 0, \\ y(t) = Cx(t), & t \geq 0 \end{cases} \quad (1)$$

with $A: \mathcal{X} \rightarrow \mathcal{X}$, $B: \mathcal{U} \rightarrow \mathcal{X}$, $C: \mathcal{X} \rightarrow \mathcal{Y}$ ($\dim(\mathcal{X}) = n$, $\dim(\mathcal{U}) = m$, $\dim(\mathcal{Y}) = p$) and denoted (A, B, C) .

In what follows we shall use the following notation: for a given map, say B , we shall denote \mathcal{B} its image. For a given map, say C , the kernel is noted as $\text{Ker}C$.

A subspace $\mathcal{V} \subset \mathcal{X}$ is said to be (A, B) -invariant if and only if $A\mathcal{V} \subset \mathcal{V} + \mathcal{B}$. We shall denote $\mathfrak{I}(A, B; S)$ the family of all the (A, B) -invariant subspaces included in the subspace $S \subset \mathcal{X}$. If $\mathcal{V} \in \mathfrak{I}(A, B; S)$, there exists at least a map $F: \mathcal{X} \rightarrow \mathcal{U}$ (called a friend of \mathcal{V}) such that $(A + BF)\mathcal{V} \subset \mathcal{V} \subset S$. The family of all friends of \mathcal{V} is written as $\mathbf{F}(\mathcal{V})$.

Let $\mathcal{V} \in \mathfrak{I}(A, B; S)$ and let us denote $\mathcal{R}_{\mathcal{V}}$ the supremal (A, B) -controllability subspace contained in \mathcal{V} . Then, the following statements are equivalent:

- $\mathcal{V} \in \mathfrak{I}(A, B; S)$ is an internally stabilizable (A, B) -invariant subspace.
- There exists a map $F \in \mathbf{F}(\mathcal{V})$ such that $\sigma(A + BF|_{\mathcal{V}}) \subset \mathbb{C}_-$, where $A + BF|_{\mathcal{V}}$ denotes the double restriction of $A + BF$ to \mathcal{V} and \mathbb{C}_- is the open left-half complex plane.
- $\sigma_{\mathcal{V}} := \sigma(\overline{A + BF})$ satisfies $\sigma_{\mathcal{V}} \subset \mathbb{C}_-$, where $\overline{A + BF}$ is the map induced in $\mathcal{V}/\mathcal{R}_{\mathcal{V}}$ by $A + BF$ ($\sigma_{\mathcal{V}}$ is independent of the choice of $F \in \mathbf{F}(\mathcal{V})$, see Property A.1 in the Appendix).

Assuming that (A, B) is stabilizable and being $\mathcal{V} \in \mathfrak{I}(A, B; S)$ an internally stabilizable (A, B) -invariant subspace, there exists at least a map $F: \mathcal{X} \rightarrow \mathcal{U}$ belonging to $\mathbf{F}(\mathcal{V})$ such that $\sigma(A + BF) \subset \mathbb{C}_-$ (see for instance [10]). In that case, we shall say that F belongs to the family $\mathbf{F}_s(\mathcal{V})$.

A set of (A, B) -invariant subspaces $\{\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_k\}$ is called compatible if there exists at least one common friend for all of them, i.e., $\bigcap_{i=1}^k \mathbf{F}(\mathcal{V}_i) \neq \emptyset$. If the family $\{\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_k\}$ is compatible, it is not difficult to show that their radical, say $\hat{\mathcal{V}}$ defined as $\hat{\mathcal{V}} := \bigcap_{i=1}^k \sum_{j \neq i} \mathcal{V}_j$, is (A, B) -invariant and $\bigcap_{i=1}^k \mathbf{F}(\mathcal{V}_i) \subset \bigcap_{i=1}^k \mathbf{F}(\mathcal{V}_i + \hat{\mathcal{V}})$ (see for instance [11, Chapter 10]).

If $\bigcap_{i=1}^k \mathbf{F}(\mathcal{V}_i) = \mathbf{F}(\hat{\mathcal{V}}) \cap \bigcap_{i=1}^k \mathbf{F}(\mathcal{V}_i + \hat{\mathcal{V}})$, then $\{\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_k\}$ is called strongly compatible [8].

III. PROBLEMS STATEMENTS

A. Decoupling

Some Definitions: First of all we shall define the Block Decoupling Problem via regular static state feedback (BDP).

Definition 1 (BDP): Given a controllable system (A, B, C) and according to a given block partition $\{C_1, C_2, \dots, C_k\}$ of C , the block decoupling problem is then defined as follows:

"Find conditions for the existence of static state feedback control laws $u(t) = Fx(t) + \sum_{i=1}^k G_i v_i(t)$ such that each block input $v_i(t)$ completely controls the block output $y_i(t) = C_i x(t)$ without affecting the $k - 1$ other block-outputs $y_j(t)$, $j \neq i$, under the constraint $G = [G_1 \ G_2 \ \dots \ G_k]$ regular, i.e., G invertible."

The number of columns in each block G_i is not fixed *a priori*. If F and G , regular, exist for the block partition $\{C_1, C_2, \dots, C_k\}$, then we shall say that (A, B, C) is a regularly block decouplable system.

Definition 2 (RBRDP): When $k = p$, BDP becomes the so-called (regular) Row By Row Decoupling Problem. Let us recall that in that case right invertibility of the system is an obvious necessary condition for RBRDP to be solvable.

Definition 3 (RBRDPS): When we add the constraint $\sigma(A + BF) \subset \mathbb{C}_-$ to RBRDP we have the so-called (regular) row by row decoupling problem with stability.

To give the solvability condition of RBRDPS in geometric terms, let us first introduce some notation. Let c_i be the i th row of C and let us denote:

- \mathcal{V}^* , respectively, \mathcal{V}_i^* , respectively, \mathcal{W}_i^* , the supremal (A, B) -invariant subspace contained in $\text{Ker}C$, respectively, $\text{Ker}c_i$, respectively, $\bigcap_{j \neq i} \text{Ker}c_j$.
- \mathcal{V}_{stab}^* , respectively, $\mathcal{V}_{i,stab}^*$, respectively, $\mathcal{W}_{i,stab}^*$, the supremal internally stabilizable (A, B) -invariant subspace contained in $\text{Ker}C$, respectively, $\text{Ker}c_i$, respectively, $\bigcap_{j \neq i} \text{Ker}c_j$.
- \mathcal{R}^* , respectively, \mathcal{R}_i^* , respectively, \mathcal{T}_i^* , the supremal (A, B) -controllability subspace contained in $\text{Ker}C$, respectively, $\text{Ker}c_i$, respectively, $\bigcap_{j \neq i} \text{Ker}c_j$.

Then we have the following theorem.

Theorem 1 [6, Theorem 3]. Assuming that (A, B) is controllable, RBRDPS is solvable if and only if

$$\begin{cases} \text{i)} & \mathcal{V}^* = \bigcap_{i=1}^p \mathcal{V}_i^* \\ \text{ii)} & \bigcap_{i=1}^p \mathcal{R}_i^* \subset \mathcal{V}_{stab}^* \end{cases}$$

Remark 1 It has to be noted that if RBRDP is solvable, $\bigcap_{i=1}^p \mathcal{R}_i^*$ is equal to the radical of $\{\mathcal{T}_1^*, \mathcal{T}_2^*, \dots, \mathcal{T}_p^*\}$ (see for instance [4, Corollary V.1]).

In an equivalent way, we have the following theorem.

Theorem 2 [6, Theorem 4]. Assuming that (A, B) is controllable, RBRDPS is solvable if and only if

$$\begin{cases} \text{i)} & \mathcal{V}^* = \bigcap_{i=1}^p \mathcal{V}_i^* \\ \text{ii)} & \mathcal{V}_{stab}^* = \bigcap_{i=1}^p \mathcal{V}_{i,stab}^* \end{cases}$$

If RBRDPS is solvable we shall say that (A, B, C) is a regularly row by row decouplable system with stability.

B. Disturbance Rejection

Let us now consider the following linear time-invariant disturbed system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + Ed(t), & t \geq 0, \\ y(t) = Cx(t), & t \geq 0 \end{cases}$$

where A, B , and C are as defined in (1) and $E: \mathcal{D} \rightarrow \mathcal{X}$ ($\dim(\mathcal{D}) = q$). The disturbance input, $d(t)$, may or may not be

measured. We shall denote this system by (A, B, C, E) .

The disturbance rejection problem with stability (DRPS) is then defined as follows.

Definition 4 (DRP-DRPS): Given a system (A, B, C, E) , find conditions for the existence of a static state feedback control law $u(t) = Fx(t) + Hd(t)$ such that $y(t)$ does not depend on the disturbance (DRP) and the closed-loop system is internally stable (DRPS).

In what follows we shall denote \mathcal{E} the image of E . The known classic geometric solvability condition of DRPS is given by the following theorem.

Theorem 3 [11, Theorem 5.8]: Assuming that (A, B) is controllable, DRPS is solvable if and only if $\mathcal{E} \subset \mathcal{V}_{stab}^*$.

Note that Theorem 3 is still valid if (A, B) is just stabilizable. If the disturbance is not measured ($H = 0$), the well-known geometric solution is $\mathcal{E} \subset \mathcal{V}_{stab}^*$.

Finally, as concerns the set where the solutions (feedbacks F) can be found, (if any), it is quite well known that, in both cases (disturbance measured or not), it is given by $\mathbf{F}_s(\mathcal{V}_{stab}^*)$.

C. Simultaneous Disturbance Rejection and Regular Row by Row Decoupling Problem With Stability (SDRDPS)

Let us consider again the disturbed linear-time invariant system (A, B, C, E)

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + Ed(t), & t \geq 0 \\ y(t) = Cx(t), & t \geq 0. \end{cases}$$

Because of decoupling requirements, we shall assume that the system without disturbance (A, B, C) is controllable and right invertible.

Definition 5 (SDRDPS): Find conditions for the existence of a static state feedback control law $u(t) = Fx(t) + Gv(t) + Hd(t)$, with $G := [G_1 \ G_2 \ \dots \ G_p]$ regular and $v(t) = [v_1^T(t) \ v_2^T(t) \ \dots \ v_p^T(t)]^T$, where $v_i(t)$ are block input vectors (since $m \geq p$) and such that $y(t)$ does not depend on the disturbance, the closed-loop system is decoupled from $v(t)$ to $y(t)$ and is internally stable.

IV. SOLUTION OF SDRDPS

For that solution, we shall characterize a family of control laws which solve SDRDPS. Before that, and in order to introduce our extension, we shall quickly recall the formal statement of the decoupling problem and the way the combined problem is solved, when no stability constraint is imposed.

To express the objective of decoupling formally, let $\{T_1, T_2, \dots, T_p\}$ be the set of subspaces which $v_i(t)$, $i \in \{1, 2, \dots, p\}$, can control. Thus T_i is the (A, B) -controllability subspace given by $T_i = (A + BF \mid B \operatorname{Im} G_i)$, $i \in \{1, 2, \dots, p\}$.

Now, the objective of decoupling can be established as follows (see for instance [11, Chapter 9]):

"Find conditions for the existence of compatible (A, B) -controllability subspaces T_i included in $\bigcap_{j \neq i} \operatorname{Ker} c_j$ and satisfying $T_i + \operatorname{Ker} c_i = \mathcal{X}$, for all $i \in \{1, 2, \dots, p\}$."

A set of controllability subspaces $\{T_1, T_2, \dots, T_p\}$ satisfying these conditions is called a solution to the row by row decoupling problem. Let us recall the following theorem.

Theorem 4 [8, Theorem 7]: A regular solution to the Row by Row Decoupling Problem exists if and only if $\mathcal{B} = \sum_{i=1}^p \mathcal{B} \cap T_i^*$. If so, $\{T_1^*, T_2^*, \dots, T_p^*\}$ is a strongly compatible solution.

A decoupling solution (if any) $F : \mathcal{X} \rightarrow \mathcal{U}$ can be picked up in the family $\bigcap_{i=1}^p \mathbf{F}(T_i^*)$. G is usually directly deduced from $\operatorname{Im} BG_i := \mathcal{B} \cap T_i^*$.

The first important result for the solvability of the simultaneous disturbance rejection and regular row by row decoupling problem

when no stability constraint is imposed (SDRDPS), is the following separation property.

Theorem 5 [1, Theorem 3]: SDRDP is solvable if and only if both RBRDP and DRP are solvable separately.

We shall here add the stability requirement. In this case, asking for the solvability condition of RBRDPS is strictly equivalent to asking for conditions of existence of a nonempty family $\bigcap_{i=1}^p \mathbf{F}_s(T_i^*) \subset \bigcap_{i=1}^p \mathbf{F}(T_i^*)$. By Theorem 1, we know that $\bigcap_{i=1}^p \mathbf{F}_s(T_i^*)$ is nonempty if and only if $\bigcap_{i=1}^p \mathcal{R}_i^* \subset \mathcal{V}_{stab}^*$.

To solve the simultaneous disturbance rejection and row by row decoupling problem with stability, we shall proceed as in the case when no stability constraint is imposed, i.e., we shall characterize a subfamily of $\bigcap_{i=1}^p \mathbf{F}_s(T_i^*)$, in terms of internally stabilizable (A, B) -invariant subspaces, in such a way that any map $F : \mathcal{X} \rightarrow \mathcal{U}$ belonging to that subfamily will also automatically reject the disturbance, provided that this latter problem is solvable.

The result established by the following theorem will play a key role in the solution of the simultaneous disturbance rejection and regular row by row decoupling problem with stability. Let $\widehat{\mathcal{W}}_{stab}^*$ denote the radical of $\{\mathcal{W}_{1\text{stab}}^*, \mathcal{W}_{2\text{stab}}^*, \dots, \mathcal{W}_{p\text{stab}}^*\}$.

Theorem 6: Let (A, B, C) be a regularly row by row decouplable system with stability. There then always exists a common stabilizing friend for the family $\{\widehat{\mathcal{W}}_{stab}^*, \widehat{\mathcal{W}}_{stab}^* + \mathcal{W}_{1\text{stab}}^*, \widehat{\mathcal{W}}_{stab}^* + \mathcal{W}_{2\text{stab}}^*, \dots, \widehat{\mathcal{W}}_{stab}^* + \mathcal{W}_{p\text{stab}}^*\}$.

The proof of this theorem will rest upon the following results, which proofs are in the Appendix.

Theorem 7: Let both $\mathcal{V}_1, \mathcal{V}_2$ be internally stabilizable (A, B) -invariant subspaces such that $\mathcal{V}_1 \subset \mathcal{V}_2$. Then, for any $F_1 \in \mathbf{F}(\mathcal{V}_1)$ such that $\sigma(A + BF_1 \mid \mathcal{V}_1) \subset \mathcal{C}_-$, there exists a map $F_2 \in \mathbf{F}(\mathcal{V}_2)$ such that $F_2 \mid \mathcal{V}_1 = F_1 \mid \mathcal{V}_1$ and $\sigma(A + BF_2 \mid \mathcal{V}_2) \subset \mathcal{C}_-$, i.e., F_2 is such that $(A + BF_2)\mathcal{V}_1 \subset \mathcal{V}_1$, $(A + BF_2)\mathcal{V}_2 \subset \mathcal{V}_2$ and $\sigma(A + BF_2 \mid \mathcal{V}_2) \subset \mathcal{C}_-$.

Lemma 1: Let (A, B, C) be a regularly row by row decouplable system with stability, then $\widehat{\mathcal{W}}_{stab}^* = \bigcap_{i=1}^p \mathcal{W}_{i\text{stab}}^* = \mathcal{V}_{stab}^*$.

Property 1: Let (A, B, C) be a regularly row by row decouplable system with stability, then $\mathcal{X} = \sum_{i=1}^p \mathcal{W}_{i\text{stab}}^*$.

Proof of Theorem 6: From Property 1 we have that the state space \mathcal{X} can be partitioned as

$$\mathcal{X} = \sum_{i=1}^p \mathcal{W}_{i\text{stab}}^* = \sum_{i=1}^p \left(\widehat{\mathcal{W}}_{stab}^* + \sum_{j=1}^i \mathcal{W}_{j\text{stab}}^* \right).$$

Note that $(\widehat{\mathcal{W}}_{stab}^* + \sum_{j=1}^i \mathcal{W}_{j\text{stab}}^*)$, for all $i \in \{1, 2, \dots, p\}$, is an internally stabilizable (A, B) -invariant subspace. Since $\widehat{\mathcal{W}}_{stab}^*$ is an internally stabilizable (A, B) -invariant subspace (indeed, we have from Lemma 1 that $\widehat{\mathcal{W}}_{stab}^* = \mathcal{V}_{stab}^*$), we can always find a map $F_0 \in \mathbf{F}(\widehat{\mathcal{W}}_{stab}^*)$ such that $\sigma(A + BF_0 \mid \widehat{\mathcal{W}}_{stab}^*) \subset \mathcal{C}_-$.

By Theorem 7, there exists a map $F_1 \in \mathbf{F}(\widehat{\mathcal{W}}_{stab}^* + \mathcal{W}_{1\text{stab}}^*)$ such that

$$\begin{aligned} F_1 \mid \widehat{\mathcal{W}}_{stab}^* &= F_0 \mid \widehat{\mathcal{W}}_{stab}^* \\ \text{and } \sigma(A + BF_1 \mid (\widehat{\mathcal{W}}_{stab}^* + \mathcal{W}_{1\text{stab}}^*)) &\subset \mathcal{C}_-. \end{aligned}$$

Applying again Theorem 7, we can find a map $F_2 \in \mathbf{F}(\widehat{\mathcal{W}}_{stab}^* + \mathcal{W}_{1\text{stab}}^* + \mathcal{W}_{2\text{stab}}^*)$ such that

$$F_2 \mid (\widehat{\mathcal{W}}_{stab}^* + \mathcal{W}_{1\text{stab}}^*) = F_1 \mid (\widehat{\mathcal{W}}_{stab}^* + \mathcal{W}_{1\text{stab}}^*)$$

and

$$\sigma(A + BF_2 \mid (\widehat{\mathcal{W}}_{stab}^* + \mathcal{W}_{1\text{stab}}^* + \mathcal{W}_{2\text{stab}}^*)) \subset \mathcal{C}_-.$$

Following the same procedure we obtain a sequence of maps

$$F_i \in \mathbf{F}\left(\widehat{\mathcal{W}}_{stab}^* + \sum_{j=1}^i \mathcal{W}_{j\text{stab}}^*\right), \quad \forall i \in \{1, 2, \dots, p\}$$

such that

$$F_i \left| \left(\widehat{\mathcal{W}}_{stab}^* + \sum_{j=1}^{i-1} \mathcal{W}_{stab}^* \right) \right. = F_{i-1} \left| \left(\widehat{\mathcal{W}}_{stab}^* + \sum_{j=1}^{i-1} \mathcal{W}_{stab}^* \right) \right.$$

and $\sigma(A + BF_i | \widehat{\mathcal{W}}_{stab}^* + \sum_{j=1}^i \mathcal{W}_{stab}^*) \subset \mathbb{C}_-$.

Finally, $F = F_p$ is the desired map.

From Lemma 1 we note that $\widehat{\mathcal{W}}_{stab}^* \subset \mathcal{W}_{stab}^*$, for all $i \in \{1, 2, \dots, p\}$. Consequently $\bigcap_{i=1}^p F(\mathcal{W}_{stab}^*) = \bigcap_{i=1}^p F(\widehat{\mathcal{W}}_{stab}^* + \mathcal{W}_{stab}^*)$.

Thus, the result established by Theorem 6 can be rewritten as follows.

Theorem 8: Let (A, B, C) be a regularly row by row decouplable system with stability, then the family $F_s(\widehat{\mathcal{W}}_{stab}^*) \cap \bigcap_{i=1}^p F_s(\mathcal{W}_{stab}^*)$ is nonempty.

Before considering the Simultaneous Regular Row By Row Decoupling and Disturbance Rejection Problem with Stability, let us first clarify the connection between the family $\bigcap_{i=1}^p F_s(\mathcal{W}_{stab}^*)$ and the family $\bigcap_{i=1}^p F_s(T_i^*)$ (proved in the Appendix).

Lemma 2: For any system (A, B, C) , the following inclusion always holds

$$F(\mathcal{W}_{stab}^*) \subset F(T_i^*), \quad \forall i \in \{1, 2, \dots, p\}.$$

An obvious corollary is the following.

Corollary 1: $F_s(\mathcal{W}_{stab}^*) \subset F_s(T_i^*)$, for all $i \in \{1, 2, \dots, p\}$.

We can now present our main result.

A. SDRDPS. The Final Result

Theorem 9: The simultaneous regular row by row decoupling and disturbance rejection problem with stability has a solution if and only if both the regular row by row decoupling problem with stability and the disturbance rejection problem with stability are solvable separately.

Proof of Theorem 9: Necessity is obvious. For sufficiency suppose that system (A, B, C) is regularly row by row decouplable with stability, then by Theorem 8, the family $F_s(\widehat{\mathcal{W}}_{stab}^*) \cap \bigcap_{i=1}^p F_s(\mathcal{W}_{stab}^*)$ is nonempty. By Corollary 1, the inclusion $F_s(\mathcal{W}_{stab}^*) \subset F_s(T_i^*)$, for all $i \in \{1, 2, \dots, p\}$, always holds and consequently

$$\bigcap_{i=1}^p F_s(\mathcal{W}_{stab}^*) \subset \bigcap_{i=1}^p F_s(T_i^*).$$

Since

$$F_s(\widehat{\mathcal{W}}_{stab}^*) \cap \bigcap_{i=1}^p F_s(\mathcal{W}_{stab}^*) \subset \bigcap_{i=1}^p F_s(\mathcal{W}_{stab}^*) \subset \bigcap_{i=1}^p F_s(T_i^*)$$

any map $F \in F_s(\widehat{\mathcal{W}}_{stab}^*) \cap \bigcap_{i=1}^p F_s(\mathcal{W}_{stab}^*)$, with a suitable map $G: \mathcal{U} \rightarrow \mathcal{U}$, will decouple, while insuring internal stability of the closed-loop system.

Now, if the disturbance rejection problem with stability is solvable, any map $F \in F_s(\widehat{\mathcal{W}}_{stab}^*) \cap \bigcap_{i=1}^p F_s(\mathcal{W}_{stab}^*)$, with (eventually, depending on the fact that $d(t)$ is measured or not) a suitable map $H: \mathcal{D} \rightarrow \mathcal{U}$, also rejects the disturbance. Indeed, from Lemma 1

we have $F(\widehat{\mathcal{W}}_{stab}^*) = F(\mathcal{V}_{stab}^*)$ and then

$$\begin{aligned} F_s(\widehat{\mathcal{W}}_{stab}^*) \cap \bigcap_{i=1}^p F_s(\mathcal{W}_{stab}^*) \\ = F_s(\mathcal{V}_{stab}^*) \cap \bigcap_{i=1}^p F_s(\mathcal{W}_{stab}^*) \subset F_s(\mathcal{V}_{stab}^*) \end{aligned}$$

which ends the proof.

V. CONCLUDING REMARKS

We have considered here the simultaneous disturbance rejection and regular row by row decoupling problem with stability (SDRDPS). We have proved that the separation property (which states that the simultaneous problem is solvable if and only if each problem is solvable separately) is still true when internal stability is required. This has been done under the non restrictive assumption that the system without disturbance is controllable. The weaker assumption of stabilizability should be sufficient, but most of the results related to decoupling which have been used here had been previously stated within this controllability assumption. Adaptation to this broader situation should be quite easy, just leading to heavier development.

Thanks to this result, it will be possible to exploit the recent geometric and structural solutions of DRPS [5, Theorem 7] and RBRDPS [6, Theorem 6] to propose a synthetic solution for SDRDPS in terms of infinite and unstable zeros.

This will be detailed in a future work.

APPENDIX

A. Proof of Theorem 7

It will rest upon the following property

Property A.1 [11, Theorem 5.7 and Corollary 5.2] Let $\mathcal{V} \in \mathcal{J}(A, B; \mathcal{X})$ be given. For any $F \in \mathbf{F}(\mathcal{V})$ write $\overline{A + BF}$ for the map induced in $\mathcal{V}/R_{\mathcal{V}}$ by $A + BF$. Then $\overline{A + BF}$ is independent of $F \in \mathbf{F}(\mathcal{V})$ and $\sigma(A + BF | \mathcal{V}) = \sigma_1 \cup \sigma_{\mathcal{V}}$, where $\sigma_1 := \sigma(A + BF | R_{\mathcal{V}})$ is freely assignable by suitable choice of $F \in \mathbf{F}(\mathcal{V})$ and $\sigma_{\mathcal{V}} := \sigma(\overline{A + BF})$ is fixed for all $F \in \mathbf{F}(\mathcal{V})$. A subspace $\mathcal{V} \in \mathcal{J}(A, B; \mathcal{X})$ is thus internally stabilizable if and only if $\sigma_{\mathcal{V}} \subset \mathbb{C}_-$.

Property A.2 [7, Lemma A.2]: Let $\mathcal{V}_1, \mathcal{V}_2 \in \mathcal{J}(A, B; \mathcal{X})$ be given. Let us denote $\overline{A + BF_2}$ the map induced by $A + BF_2$ in $\mathcal{V}_2/\mathcal{V}_1$. If $\mathcal{V}_1 \subset \mathcal{V}_2$ then for any map $F_1 \in \mathbf{F}(\mathcal{V}_1)$, there exists a map $F_2 \in \mathbf{F}(\mathcal{V}_2)$ such that $F_2 | \mathcal{V}_1 = F_1 | \mathcal{V}_1$. Moreover, if $\mathcal{V}_2 = \mathcal{V}_1 + \mathcal{R}_{\mathcal{V}_2}$, then $\sigma(\overline{A + BF_2})$ is freely assignable by suitable choice of $F_2 \in \mathbf{F}(\mathcal{V}_2)$ and $\sigma_{\mathcal{V}_2} \subset \sigma_{\mathcal{V}_1}$.

Proof of Theorem 7: Let $F_1 \in \mathbf{F}(\mathcal{V}_1)$ be a map such that $\sigma(A + BF_1 | \mathcal{V}_1) \subset \mathbb{C}_-$. By Property A.2, we can choose a map $F_c \in \mathbf{F}(\mathcal{L} := \mathcal{V}_1 + \mathcal{R}_{\mathcal{V}_2})$ such that $F_c | \mathcal{V}_1 = F_1 | \mathcal{V}_1$ and $\sigma(\overline{A + BF_c}) \subset \mathbb{C}_-$, where $\overline{A + BF_c}$ is the map induced by $A + BF_c$ in $\mathcal{L}/\mathcal{V}_1$. Applying again Property A.2, we can choose a map $F_2 \in \mathbf{F}(\mathcal{V}_2)$ such that $F_2 | \mathcal{L} = F_c | \mathcal{L}$.

Let us now denote $\overline{A + BF_2}$ the map induced by $A + BF_2$ in $\mathcal{V}_2/\mathcal{L}$. Note that $\sigma(A + BF_2 | \mathcal{V}_2) = \sigma(A + BF_1 | \mathcal{V}_1) \cup \sigma(\overline{A + BF_c}) \cup \sigma(\overline{A + BF_2})$.

Since the internal stabilizability of the subspace \mathcal{V}_2 means that $\sigma_{\mathcal{V}_2} \subset \mathbb{C}_-$ (see Property A.1), and since $\sigma(\overline{A + BF_c}) \subset \sigma_{\mathcal{V}_2}$, we obtain $\sigma(A + BF_2 | \mathcal{V}_2) \subset \mathbb{C}_-$, which ends the proof.

B. Proof of Lemma 1

To prove the result established in Lemma 1 we shall use some nice properties of regularly row by row decouplable systems with stability.

Property B.1: Let (A, B, C) be a regularly row by row decouplable system with stability, then $\mathcal{V}_{i \text{ stab}}^* = \sum_{j \neq i} \mathcal{W}_j^* \text{ stab}$, for all $i \in \{1, 2, \dots, p\}$.

Now, to prove this, we shall need some properties of controllable regularly block decouplable systems. Let us first complete our previous notation.

Let \bar{C}_i be the C matrix without the i th block, noted C_i , and let us denote:

- $\mathcal{R}_i^*(\text{Ker } \bar{C}_i)$, respectively, $\mathcal{T}_i^*(\text{Ker } \bar{C}_i)$, the supremal (A, B) -controllability subspace contained in $\text{Ker } C_i$, respectively, $\text{Ker } \bar{C}_i$.

Property B.2 [4, Properties V.1 and V.3]: Let (A, B, C) be a controllable regularly block decouplable system, then

$$\sum_{i=1}^k \mathcal{T}_i^*(\text{Ker } \bar{C}_i) = \mathcal{X} \text{ and } \mathcal{R}_i^*(\text{Ker } C_i) = \sum_{j \neq i} \mathcal{T}_j^*(\text{Ker } \bar{C}_j), \quad \forall i \in \{1, 2, \dots, k\}.$$

Property B.2 let us write the following.

Property B.3: Let (A, B, C) be a controllable regularly row by row decouplable system. For any $j \in \mathbf{l}$ and any $z \in \mathbf{r}$, where \mathbf{l} and \mathbf{r} are any sets contained in $\{1, 2, \dots, p\}$ such that $\mathbf{l} \cap \mathbf{r} = \emptyset$, we have

$$\bigcap_{j \in \mathbf{l}} \mathcal{V}_j^* \text{ stab} + \bigcap_{z \in \mathbf{r}} \mathcal{V}_z^* \text{ stab} = \mathcal{X}.$$

Proof of Property B.3: Build a block, say C_{H_j} , formed with the rows of C which index belongs to \mathbf{l} and build a block, say C_{H_z} , formed with the rows of C which index belongs to \mathbf{r} . Let us now denote the supremal (A, B) -controllability subspaces contained in $\text{Ker } C_{H_j}$ and $\text{Ker } C_{H_z}$ by $\mathcal{R}_{H_j}^*(\text{Ker } C_{H_j})$ and $\mathcal{R}_{H_z}^*(\text{Ker } C_{H_z})$, respectively. From Property B.2 we have $\mathcal{R}_{H_j}^*(\text{Ker } C_{H_j}) + \mathcal{R}_{H_z}^*(\text{Ker } C_{H_z}) = \mathcal{X}$. Since $\mathcal{R}_{H_j}^*(\text{Ker } C_{H_j}) \subset \bigcap_{j \in \mathbf{l}} \mathcal{V}_j^* \text{ stab}$ and $\mathcal{R}_{H_z}^*(\text{Ker } C_{H_z}) \subset \bigcap_{z \in \mathbf{r}} \mathcal{V}_z^* \text{ stab}$, we get $\bigcap_{j \in \mathbf{l}} \mathcal{V}_j^* \text{ stab} + \bigcap_{z \in \mathbf{r}} \mathcal{V}_z^* \text{ stab} = \mathcal{X}$.

Proof of Property B.1: Note that Property B.1 obviously holds for $p = 2$. Let us give the proof for $p = 3$.

Since (A, B, C) is regularly row by row decouplable with stability, the same obviously holds for any selection of outputs within C , namely for (A, B, \bar{C}_i) , for all $i \in \{1, 2, \dots, p\}$. Theorem 2 thus implies that $\mathcal{W}_j^* \text{ stab} = \bigcap_{k \neq j} \mathcal{V}_k^* \text{ stab}$ for any $j = 1, 2, \dots, p$. Then

$$\begin{aligned} \sum_{j \neq 1} \mathcal{W}_j^* \text{ stab} &= \mathcal{W}_2^* \text{ stab} + \mathcal{W}_3^* \text{ stab} \\ &= \mathcal{V}_1^* \text{ stab} \cap \mathcal{V}_3^* \text{ stab} + \mathcal{V}_1^* \text{ stab} \cap \mathcal{V}_2^* \text{ stab}. \end{aligned}$$

From Property B.3 we have that $\mathcal{V}_3^* \text{ stab} + \mathcal{V}_1^* \text{ stab} \cap \mathcal{V}_2^* \text{ stab} = \mathcal{X}$ and consequently $\mathcal{V}_1^* \text{ stab} \cap (\mathcal{V}_3^* \text{ stab} + \mathcal{V}_1^* \text{ stab} \cap \mathcal{V}_2^* \text{ stab}) = \mathcal{V}_1^* \text{ stab} \cap \mathcal{V}_1^* \text{ stab} + \mathcal{V}_1^* \text{ stab} \cap \mathcal{V}_2^* \text{ stab} = \mathcal{V}_1^* \text{ stab} \cap \mathcal{X} = \mathcal{V}_1^* \text{ stab}$. From this follows that $\sum_{j \neq 1} \mathcal{W}_j^* \text{ stab} = \mathcal{V}_1^* \text{ stab}$.

In a similar way we can show that $\sum_{j \neq 2} \mathcal{W}_j^* \text{ stab} = \mathcal{V}_2^* \text{ stab}$ and $\sum_{j \neq 3} \mathcal{W}_j^* \text{ stab} = \mathcal{V}_3^* \text{ stab}$.

A similar treatment can be applied for the case $p \geq 3$. \square

We can now achieve the proof of Lemma 1

Since (A, B, C) is regularly row by row decouplable with stability it is easy to see from Theorem 2 that $\mathcal{W}_i^* \text{ stab} = \bigcap_{j \neq i} \mathcal{V}_j^* \text{ stab}$, for all $i \in \{1, 2, \dots, p\}$, and consequently $\bigcap_{i=1}^p \mathcal{W}_i^* \text{ stab} = \bigcap_{i=1}^p \bigcap_{j \neq i} \mathcal{V}_j^* \text{ stab} = \bigcap_{i=1}^p \mathcal{V}_i^* \text{ stab} =: \mathcal{V}_{\text{stab}}^*$. From this, Property B.1 and Theorem 2 we have

$$\widehat{\mathcal{W}_{\text{stab}}^*} := \bigcap_{i=1}^p \sum_{j \neq i} \mathcal{W}_j^* \text{ stab} = \bigcap_{i=1}^p \mathcal{V}_i^* \text{ stab} = \mathcal{V}_{\text{stab}}^* = \bigcap_{i=1}^p \mathcal{W}_i^* \text{ stab}. \quad \square$$

C. Proof of Property 1

Direct from Property B.2. Indeed

$$\mathcal{X} = \sum_{i=1}^p \mathcal{T}_i^* \subset \sum_{i=1}^p \mathcal{W}_i^* \text{ stab}.$$

D. Proof of Lemma 2

Lemma 2 is direct from the following results.

Theorem D.1: Let $\mathcal{V}_{\text{stab}}^*(S)$ be the supremal internally stabilizable (A, B) -invariant subspace contained in S and let $\mathcal{R}^*(S)$ be the supremal (A, B) -controllability subspace contained in S . For any $F \in \mathbf{F}(\mathcal{V}_{\text{stab}}^*(S))$ we have

$$\mathcal{R}^*(S) = \langle A + BF \mid \mathcal{B} \cap \mathcal{V}_{\text{stab}}^*(S) \rangle.$$

Proof of Theorem D.1: Let $\mathcal{V}^*(S)$ be the supremal (A, B) -invariant subspace contained in S . Note that

$$\mathcal{B} \cap \mathcal{V}_{\text{stab}}^*(S) = \mathcal{B} \cap \mathcal{V}^*(S) = \mathcal{B} \cap \mathcal{R}^*(S).$$

Indeed, $\mathcal{B} \cap \mathcal{R}^*(S) \subset \mathcal{B} \cap \mathcal{V}_{\text{stab}}^*(S) \subset \mathcal{B} \cap \mathcal{V}^*(S) = \mathcal{B} \cap \mathcal{R}^*(S)$ (see [11, Theorem 5.5]). With $F \in \mathbf{F}(\mathcal{V}_{\text{stab}}^*(S))$, write $\mathcal{R} := \langle A + BF \mid \mathcal{B} \cap \mathcal{V}_{\text{stab}}^*(S) \rangle$.

Property A.2 states the existence of a map $F' \in \mathbf{F}(\mathcal{V}^*(S))$ such that $F \mid \mathcal{V}_{\text{stab}}^*(S) = F' \mid \mathcal{V}_{\text{stab}}^*(S)$. Then $\mathcal{R} = \langle A + BF' \mid \mathcal{B} \cap \mathcal{V}_{\text{stab}}^*(S) \rangle = \langle A + BF' \mid \mathcal{B} \cap \mathcal{V}^*(S) \rangle =: \mathcal{R}^*(S)$, since $\mathcal{R}^*(S) = \langle A + BF' \mid \mathcal{B} \cap \mathcal{V}^*(S) \rangle$, for any $F' \in \mathbf{F}(\mathcal{V}^*(S))$ [11, Theorem 5.5], which ends the proof.

Corollary D.1:

$$\mathbf{F}(\mathcal{V}_{\text{stab}}^*(S)) \subseteq \mathbf{F}(\mathcal{R}^*(S)).$$

REFERENCES

- [1] M. Chang and I. B. Rhodes, "Disturbance localization in linear systems with simultaneous decoupling, pole assignment, or stabilization," *IEEE Trans. Automat. Contr.*, vol. AC-20, pp. 518–523, Aug. 1975.
- [2] J. M. Dion, C. Commault, and J. Montoya, "Simultaneous decoupling and disturbance rejection: A structural approach," *Int. J. Contr.*, vol. 59, no. 5, pp. 1325–1344, 1994.
- [3] E. Fabian and W. M. Wonham, "Decoupling and disturbance rejection," *IEEE Trans. Automat. Contr.*, vol. AC-20, pp. 399–401, Mar. 1975.
- [4] S. Icart, J. F. Lafay and M. Malabre, "A unified study of the fixed modes of systems decoupled via regular static state feedback," in *Proc. Joint Conf. New Trends in System Theory*, Genova, Italy, 1990, pp. 425–432.
- [5] M. Malabre and J. C. Martínez García, "The modified disturbance rejection problem with stability: a structural approach," in *Proc. Second European Contr. Conf. (ECC '93)*, Groningen, The Netherlands, 1993, pp. 1119–1124.
- [6] J. C. Martínez García and M. Malabre, "The row by row decoupling problem with stability: A structural approach," *IEEE Trans. Automat. Control*, to appear, 1995.
- [7] A. S. Morse, "Structure and design of linear model following systems," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 346–354, Aug. 1973.
- [8] A. S. Morse and W. M. Wonham, "Status of noninteracting control," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 568–581, Dec. 1971.
- [9] P. N. Paraskevopoulos, F. N. Koumboulis, and K. G. Tzlerakis, "Disturbance rejection with simultaneous decoupling of linear time invariant systems," in *Proc. First European Contr. Conf. (ECC '91)*, Grenoble, France, 1991, pp. 1784–1788.
- [10] J. M. Schumacher, "A complement on pole placement," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 281–282, Apr. 1980.
- [11] W. Wonham, *Linear Multivariable Control: A Geometric Approach*. New York: Springer-Verlag, 1985.

Robust Controller Design for Delay Systems in the Gap-Metric

Akira Kojima and Shintaro Ishijima

Abstract—A robust stabilization problem in the gap-metric is discussed for a system with delays in control. We derive a design procedure of robust controllers based on finite-dimensional Riccati equations and it is shown that the resulting controller has a structure of observer-based predictive action. By employing completing the square argument, a game theoretic interpretation is also provided on the trade-off between the initial-uncertainties and attenuating the disturbance caused by plant uncertainties.

I. INTRODUCTION

In recent years, a robust stabilization problem in the gap-metric has received much attention [7], [17] and, for finite-dimensional systems, it is recognized that the design procedure enables us to take a trade-off between the robustness and the performance of closed-loop system [18]. Moreover, this argument has a merit to obtain a control law in a closed formula, which has a well-known feature of LQG-controllers.

For a class of infinite-dimensional systems, the robust stabilization problem in the gap-metric has been studied by various authors [3], [8], [19]. In [3], the maximum robustness margin and the admissible controllers are derived based on an input/output argument; alternatively an operator theory derivation is applied in [8] and provides a number of results on the gap and graph metrics. Especially for a system with delays in control: $\Sigma_P(s) = R(s) \cdot e^{-sh}$, a closed formula is given for the evaluation of maximum robustness margin [19].

Even for the time delay systems, however, the design procedure of robust controller is not solved in the general setting. In [8], the procedure is derived for single-input/single-output (SISO) first-order input delay systems and, in [19], a sequence of finite-dimensional controllers, which approach optimality, is given for SISO retarded delay systems. Very recently in the preparation of this article, alternative design procedure is reported for SISO input delay systems [6] associated with [8]. It should be also noted that, even if we employ H^∞ optimization method for the infinite-dimensional systems [11], [24], the computational work needed to check the solvability involves quite enormous calculations.

In this paper, we focus on the following class of input delays systems

$$\begin{aligned}\Sigma_P: \quad & \dot{x}(t) = Ax(t) + Bu(t-h) \\ & y(t) = Cx(t), \\ & u(t) \in R^m, x(t) \in R^n, y(t) \in R^r\end{aligned}$$

and consider the robust stabilization problem in the gap-metric. Our objective in this paper is to derive a design procedure of robust controllers based on finite-dimensional Riccati equations. The derived control law has a structure of observer-based predictive action

$$\begin{aligned}u(t) &= K_0 \hat{x}(t) + \int_{-h}^0 K_1(\beta) u(t+\beta) d\beta \\ \dot{\hat{x}}(t) &= A \hat{x}(t) + Bu(t-h) + L(y(t) - C \hat{x}(t))\end{aligned}$$

Manuscript received January 4, 1994.

The authors are with the Department of Electronic Systems Engineering, Tokyo Metropolitan Institute of Technology, Asahigaoka 6-6, Hino-city, Tokyo 191, Japan.

IEEE Log Number 9407225.

and the procedure does not require the restrictions on the degree of plant and the number of input/output channels. A game theoretic interpretation is also provided on the trade-off between the initial-uncertainties and attenuating the disturbance caused by plant uncertainties.

In the following, we first reformulate the robust stabilization problem as an H^∞ output feedback problem for time delay systems. Then applying completing the square argument of particular quadratic forms [22], [13], [2], we directly derive a robust control law based on auxiliary operator type Riccati equations. The approach employed here also enables us to clarify the corresponding worst-case disturbance and, further, provides game-theoretic interpretations on the trade-off between attenuating the disturbance and the initial uncertainties. Finally, we show that the solutions to the operator-type Riccati equations are constructively given based on finite-dimensional Riccati equations and summarize the design procedure.

II. FORMULATION AND PRELIMINARIES

For a system with delays in control

$$\begin{aligned}\Sigma_P: \quad & \dot{x}(t) = Ax(t) + Bu(t-h) \\ & y(t) = Cx(t), \\ & u(t) \in R^m, x(t) \in R^n, y(t) \in R^r\end{aligned}\tag{1}$$

let $\Sigma_P(s) := R(s) \cdot e^{-sh} = \tilde{M}^{-1}(\tilde{N} \cdot e^{-sh})$ and $[\tilde{M}, \tilde{N}]$ be the normalized left coprime factorization of the rational part $R(s)$. Then it is clarified by [19] that the maximum robustness margin is analytically given in the following way.

Fact 1 (Partington et al. [19]): Let $[\tilde{N}, \tilde{M}] := [0, I] + C_H(sI - A_H)^{-1} [B_H, H_H]$ and σ_1 be the largest solution to the equation

$$\det \left\{ \begin{aligned} & [-\sigma^{-1} Q_H, I] \\ & \cdot \exp \left(\begin{bmatrix} A_H & \sigma^{-1} B_H B_H^T \\ -\sigma^{-1} C_H^T C_H & -A_H^T \end{bmatrix} \cdot h \right) \\ & \cdot \begin{bmatrix} \sigma^{-1} P_H \\ I \end{bmatrix} \end{aligned} \right\} = 0$$

where P_H and Q_H satisfy the Lyapunov equations respectively

$$\begin{aligned}A_H P_H + P_H A_H^T + B_H B_H^T + H_H H_H^T &= 0, \\ A_H^T Q_H + Q_H A_H + C_H^T C_H &= 0.\end{aligned}$$

Then the maximum robustness margin in the gap-metric is given by $\epsilon_{opt} = \sqrt{1 - \sigma_1^2}$.

The analytic result stated above provides deep insight on the effect of time delay h and, further, promises the first stage to attack the robust controller design problem in the gap-metric. Our objective in this paper is to derive a design procedure of robust controllers in the framework of finite-dimensional operations.

Robust Controller Design Problem

For the time delay system Σ_P , derive a design procedure of robust controllers such that the closed-loop system is stabilized against the perturbations $\{\Delta := \Delta = [\Delta_M, \Delta_N]\}$

$$\begin{aligned}\Sigma_P^\Delta(s) &= (\tilde{M} + \Delta_M)^{-1} (\tilde{N} \cdot e^{-sh} + \Delta_N) \\ \{\Delta &= [\Delta_M, \Delta_N]; \Delta \in H^\infty, \|\Delta\|_\infty < \epsilon\}, \\ &\epsilon < \epsilon_{opt} < 1\end{aligned}\tag{2}$$

where $[\tilde{M}, \tilde{N}]$ are the normalized left coprime factors of the rational part $R(s)$ and $\{\Sigma_P^\Delta\}$ indicates the plant set with the radius ϵ in the H^∞ gap-metric.

On the realization of nominal plant Σ_P (1), we make a following assumption:

A1) (C, A, B) is observable and controllable.

The robust stabilization problem stated here is well posed as an H^∞ optimization problem to find a stabilizing law $K(s)$ which satisfies the following condition

$$\left\| \begin{bmatrix} I \\ K \end{bmatrix} (I - \Sigma_P K)^{-1} \hat{M}^{-1} \right\|_\infty < \epsilon^{-1}. \quad (3)$$

Moreover condition (3) is equivalently reformulated as follows

$$\left\| \begin{bmatrix} I \\ K \end{bmatrix} (I - \Sigma_P K)^{-1} \hat{M}^{-1} [\hat{M}, \hat{N} \cdot e^{-sh}] \right\|_\infty = \left\| \begin{bmatrix} I \\ K \end{bmatrix} (I - \Sigma_P K)^{-1} [I, R \cdot e^{-sh}] \right\|_\infty < \epsilon^{-1} \quad (4)$$

where the delay term e^{-sh} in the left hand side is artificially introduced in order to finally reduce the design procedure into finite-dimensional operations.

In the sequel, we define an augmented time delay system associated with (4)

$$\begin{aligned} \Sigma: \quad \dot{x}(t) &= Ax(t) + Bw_2(t-h) + Bu(t-h) \\ z_1(t) &= Cx(t) + w_1(t) \\ z_2(t) &= u(t) \\ y(t) &= Cx(t) + w_1(t) \end{aligned} \quad (5)$$

and discuss an H^∞ output feedback problem for the system Σ . Namely, the problem is to find an admissible output feedback law such that:

- C1) the closed-loop system is internally stable, and
- C2) the closed-loop transfer function from the disturbance $w := [w_1^T, w_2^T]^T$ to the regulated output $z := [z_1^T, z_2^T]^T: \Sigma_{zw}(s)$ satisfies $\|\Sigma_{zw}\|_\infty < \gamma$ for a given constant $\gamma := \epsilon^{-1}$, $\epsilon < \epsilon_{opt} < 1$.

By employing a state-space setup investigated in [10], [20], [21], we next describe the time delay system Σ on appropriately defined Hilbert space. On a Hilbert space $\mathcal{X} := R^n \times L_2(-h, 0; R^m)$ endowed with the inner product

$$\begin{aligned} \langle \psi, \phi \rangle &:= \psi^0 T \phi^0 + \int_{-h}^0 \psi^{1T}(\beta) \phi^1(\beta) d\beta, \\ \psi &= \begin{bmatrix} \psi^0 \\ \psi^1 \end{bmatrix} \in \mathcal{X}, \quad \phi = \begin{bmatrix} \phi^0 \\ \phi^1 \end{bmatrix} \in \mathcal{X} \end{aligned} \quad (6)$$

the augmented system Σ can be written in a form of an evolution equation [10], [20], [21]

$$\begin{aligned} \Sigma: \quad \dot{\hat{x}}(t) &= \mathcal{A}\hat{x}(t) + \mathcal{B}w_2(t) + \mathcal{B}u(t) \\ z_1(t) &= \mathcal{C}\hat{x}(t) + w_1(t) \\ z_2(t) &= u(t) \\ y(t) &= \mathcal{C}\hat{x}(t) + w_1(t). \end{aligned} \quad (7)$$

The operator \mathcal{A} is an infinitesimal generator defined as follows

$$\mathcal{A}\phi = \begin{bmatrix} A\phi^0 + B\phi^1(-h) \\ \phi^{1'} \end{bmatrix}, \quad \mathcal{D}(\mathcal{A}) = \{\phi \in \mathcal{X} : \phi^1 \in W^{1,2}(-h, 0; R^m), \phi^1(0) = 0\} \quad (8)$$

where $W^{1,2}$ denotes the Sobolev space of R^m -valued, absolutely continuous functions with square integrable derivatives on $[-h, 0]$. The adjoint operator \mathcal{A}^* is given by

$$\begin{aligned} \mathcal{A}^* \psi &= \begin{bmatrix} A^T \psi^0 \\ -\psi^{1'} \end{bmatrix}, \\ \mathcal{D}(\mathcal{A}^*) &= \{\psi \in \mathcal{X} : \psi^1 \in W^{1,2}(-h, 0; R^m), \psi^1(-h) = B^T \psi^0\} \end{aligned} \quad (9)$$

and, extending the state space \mathcal{X} to $\mathcal{V} := \mathcal{D}(\mathcal{A}^*)^*$, it is verified that $\mathcal{D}_{\mathcal{V}}(\mathcal{A}) = \mathcal{X}$ and the separate Hilbert space \mathcal{V}^* , \mathcal{X} and \mathcal{V} are with continuous, dense injections satisfying

$$\mathcal{V}^* \subset \mathcal{X} \subset \mathcal{V} \quad (\mathcal{X} = \mathcal{X}^*).$$

The input/output operators \mathcal{B} and \mathcal{C} are defined as follows

$$\begin{aligned} \mathcal{B}: R^m &\rightarrow \mathcal{V}, \quad \mathcal{B}^* \psi = \psi^1(0) \quad (\psi \in \mathcal{V}^* = \mathcal{D}(\mathcal{A}^*)) \\ \mathcal{C}: \mathcal{X} &\rightarrow R^r, \quad \mathcal{C}\phi = C\phi^0 \quad (\phi \in \mathcal{X}). \end{aligned} \quad (10)$$

III. ROBUST CONTROLLER DESIGN

In this section, we first construct an admissible control law based on auxiliary operator type Riccati equations. By employing completing the square argument of particular quadratic forms [22], [13], [2], the control law is directly derived together with the corresponding worst-case disturbance. Then it is clarified that the solution to the operator Riccati equations are constructively given based on finite-dimensional Riccati equations.

The admissible control law for the posed H^∞ output feedback problem is formally characterized as follows.

Theorem 2: Let $P > 0$ ($P \in \mathcal{L}(\mathcal{X})$) and $S > 0$ ($S \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$) be the stabilizing solutions to the operator Riccati equations

$$\begin{aligned} P\mathcal{A}^* \psi + \mathcal{A}P\psi - PC^*CP\psi + \mathcal{B}\mathcal{B}^* \psi &= 0, \quad \psi \in \mathcal{V}^* \quad (11) \\ S\left(\mathcal{A} + \frac{1}{\gamma^2 - 1} \cdot PC^*C\right)\phi &+ \left(\mathcal{A} + \frac{1}{\gamma^2 - 1} \cdot PC^*C\right)^* \cdot S\phi - S\mathcal{B}\mathcal{B}^*S\phi \\ &+ \frac{1}{\gamma^2 - 1} \cdot SPC^*CPS\phi + \frac{\gamma^2}{\gamma^2 - 1} C^*C\phi = 0, \end{aligned} \quad \phi \in \mathcal{X} \quad (12)$$

then an admissible control law satisfying (C1) and (C2) is given by

$$\begin{aligned} u(t) &= -\mathcal{B}^* S \hat{x}(t) \\ \dot{\hat{x}}(t) &= \mathcal{A}\hat{x}(t) + \mathcal{B}u(t) + PC^*(y(t) - \mathcal{C}\hat{x}(t)), \\ \hat{x}(0) &= 0. \end{aligned} \quad (13)$$

In saying the stabilizing solutions P and S , we mean that the operators $\mathcal{A} - PC^*C$, $\mathcal{A} - \mathcal{B}\mathcal{B}^*S + (1/(\gamma^2 - 1)) \cdot PC^*C(I + PS)$ generate exponentially stable semigroups.

Remark 1: For the linear operator $S \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$, the positive definiteness $S > 0$ requires that $S = S^* \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$ and, for given $\phi \neq 0$ ($\phi \in \mathcal{V}$), $\langle \phi, S\phi \rangle_{\mathcal{V}, \mathcal{V}^*} > 0$ hold. The properties of Hilbert adjoint operations between \mathcal{V} and \mathcal{V}^* are found in [24, Section 2.5].

Before describing the complete proof, we give here some preliminaries and motivations. The Riccati equation (11) is with the unbounded operator $\mathcal{B}\mathcal{B}^*$ and the structure of solution is not generally clarified [20]. However, in case we deal with the augmented system Σ , the solution $P > 0$ ($P \in \mathcal{L}(\mathcal{X})$) is constructively given based on finite-dimensional Riccati equation.

Lemma 3: Let $P > 0$ be the solution to the finite-dimensional Riccati equation

$$PA^T + AP - PC^T CP + BB^T = 0 \quad (14)$$

then the stabilizing solution $P > 0$ ($P \in \mathcal{L}(\mathcal{X})$) to (11) is given by

$$P = \begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix} \in \mathcal{L}(\mathcal{X}) \quad (15)$$

and there exists a bounded inverse on \mathcal{X} .

Proof Direct verified by substituting the solution (15) to (11) Q.E.D.

The above lemma guarantees the existence of bounded inverse $P^{-1} \in \mathcal{L}(\mathcal{V})$ and enables us to directly apply completing the square argument associated with the operators S and P^{-1} [22]–[23].

Proof of Theorem 2 Let $\epsilon = \tau - \bar{\tau}$ and describe the closed loop system with respect to the state \underline{x}, ϵ

$$\begin{aligned} \Sigma: \quad \begin{bmatrix} \underline{x}(t) \\ \epsilon(t) \end{bmatrix} &= \begin{bmatrix} \mathcal{A} - \mathcal{B}\mathcal{K}^*S & \mathcal{P}\mathcal{C}^*\mathcal{C} \\ 0 & \mathcal{A} - \mathcal{P}\mathcal{C}^*\mathcal{C} \end{bmatrix} \begin{bmatrix} \underline{x}(t) \\ \epsilon(t) \end{bmatrix} \\ &+ \begin{bmatrix} \mathcal{P}\mathcal{C}^* & 0 \\ -\mathcal{P}\mathcal{C}^*\mathcal{B} & 0 \end{bmatrix} u(t) \\ \dot{\epsilon}(t) &= \begin{bmatrix} \mathcal{C} & \mathcal{C} \\ -\mathcal{B}^*S & 0 \end{bmatrix} \begin{bmatrix} \underline{x}(t) \\ \epsilon(t) \end{bmatrix} \\ &+ \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} u(t) \\ u(t) &= \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}, \quad \dot{\epsilon}(t) = \begin{bmatrix} \dot{\epsilon}_1(t) \\ \dot{\epsilon}_2(t) \end{bmatrix} \quad (16) \end{aligned}$$

In the following we will show that the closed loop system Σ satisfies the conditions C1) and C2).

Internal Stability C1) Since $S > 0$ and $P > 0$ are the stabilizing solutions, it is easily verified that the operators $\mathcal{A} - \mathcal{B}\mathcal{K}^*S$ and $\mathcal{A} - \mathcal{P}\mathcal{C}^*\mathcal{C}$ generate exponentially stable semigroups. While the operator $\mathcal{P}\mathcal{C}^*\mathcal{C}$ is bounded. Hence it follows that the closed loop system is internally stable [12].

Disturbance Attenuation C2) To verify the inequality $\|\Sigma\|_{\infty} \leq \gamma$, i.e. $\|\dot{\epsilon}\|_2 \leq \gamma \|\dot{u}\|_2$, $u \neq 0$ we first focus on the restricted class of disturbances

$$u \in C^1([0, \infty), R^{+}) \cap I([0, \infty), R^{+}) \quad (17)$$

which is continuously differentiable and denote the disturbances as $u \in C^1 I([0, \infty), R^{+})$. The space $C^1 I([0, \infty), R^{+})$ is dense in $I([0, \infty), R^{+})$ and it is known that if $\underline{x}(0), \epsilon(0) \in \mathcal{D}(\mathcal{A})$ and $u \in C^1 I([0, \infty), R^{+})$ the solution $[\underline{x}(t), \epsilon(t)]$ ($t > 0$) is everywhere differentiable [4]. Consider the functional

$$V(t) = \langle \underline{x}(t), S\underline{x}(t) \rangle + \langle \epsilon(t), \mathcal{P}^{-1}\epsilon(t) \rangle$$

on V then differentiating both sides with respect to t inserting (11) and (12) we have

$$\begin{aligned} \dot{V}(t) &= -\|\dot{\epsilon}\|_2^2 + \gamma^2 \|\dot{u}\|_2^2 - \langle \dot{\epsilon}, -\mathcal{B}^*S\mathcal{P}^{-1}\epsilon \rangle \\ &- (\gamma^2 - 1) \left\| u_1 + \mathcal{C}\epsilon - \frac{1}{\gamma^2 - 1} \mathcal{C}(I + \mathcal{P}S)\underline{x} \right\|_2^2 \end{aligned}$$

Integrating both sides with respect to t over the interval $[0, \infty)$ we finally obtain

$$\begin{aligned} \langle \underline{x}(\infty), S\underline{x}(\infty) \rangle + \langle \epsilon(\infty), \mathcal{P}^{-1}\epsilon(\infty) \rangle &- \langle \underline{x}(0), S\underline{x}(0) \rangle - \langle \epsilon(0), \mathcal{P}^{-1}\epsilon(0) \rangle \\ &= -\|\dot{\epsilon}\|_2^2 + \gamma^2 \|\dot{u}\|_2^2 - \int_0^\infty \left\| u_1 + \mathcal{C}\epsilon - \frac{1}{\gamma^2 - 1} \mathcal{C}(I + \mathcal{P}S)\underline{x} \right\|_2^2 dt \quad (18) \end{aligned}$$

Since Σ is shown to be internally stable, $\underline{x}(\infty) = \epsilon(\infty) = 0$ and without loss of generality in the input-output analysis we can assume the initial state as $\underline{x}(0) = \epsilon(0) = 0$. Hence it follows from the equality (18) that

$$\|\dot{\epsilon}\|_2 \leq \gamma \|\dot{u}\|_2, \quad u \neq 0 \quad (19)$$

holds for the continuously differentiable disturbances $u \in C^1 I([0, \infty), R^{+})$, $u \neq 0$.

For given $u \in L_2([0, \infty), R^{+})$, there exists a sequence $\{u_n\} \subset C^1 I([0, \infty), R^{+})$ which converges to $u \in L_2([0, \infty), R^{+})$. Hence it then follows from the fact that $\underline{x}(t), \epsilon(t)$ ($t > 0$) continuously depends on $\{u_n\}$ that (19) holds for all the disturbances, $u \in I([0, \infty), R^{+})$, $u \neq 0$. Q.E.D.

Remark 2 Equality (18) clarifies the worst-case disturbance in the H^∞ output feedback setting. The worst-case disturbance $u = u^0$ is given by

$$\begin{aligned} u^0 &= \begin{bmatrix} u_1^0 \\ u_2^0 \end{bmatrix}, \quad u_1^0 = -\mathcal{C}\epsilon + \frac{1}{\gamma^2 - 1} \mathcal{C}(I + \mathcal{P}S)\underline{x} \\ u_2^0 &= \mathcal{B}^*S\mathcal{P}^{-1}\epsilon \end{aligned}$$

and with the control (13) say $u = u^0$ the strategies (u^0, u^0) form a saddle point as follows [22]–[13]

$$\begin{aligned} J(u, u) &= \|\dot{\epsilon}\|_2^2 - \gamma^2 \|\dot{u}\|_2^2 \\ J(u^0, u) &\leq J(u^0, u^0) \leq J(u, u^0) \end{aligned}$$

Remark 3 The completing the square argument employed here provide direct interpretations on the trade off between attenuating the disturbance and the initial uncertainties.

Let $\epsilon(0) = \epsilon_0$, $u(\tau) = u_\epsilon(\tau)$ ($-h \leq \tau < 0$) be the initial uncertainties of plant and apply completing the square argument in like fashion. Then we have the following inequality on the disturbance and the initial uncertainty

$$\begin{aligned} \|\dot{\epsilon}\|_2 &\leq \|\dot{u}\|_2 + \left(\epsilon_0^T \mathcal{P}^{-1} \epsilon_0 + \int_{-h}^0 u^T(\tau) u_\epsilon(\tau) d\tau \right)^{1/2} \\ \text{i.e.} \quad \frac{\|\dot{\epsilon}\|_2}{\|\dot{u}\|_2 + \left(\epsilon_0^T \mathcal{P}^{-1} \epsilon_0 + \int_{-h}^0 u^T(\tau) u_\epsilon(\tau) d\tau \right)^{1/2}} &\leq \gamma \quad (20) \end{aligned}$$

In inequality (20) the solution $P > 0$ to the filtering Riccati equation (14) plays a part of weighting matrix between the uncertainty caused by disturbance u and the initial plant uncertainty ϵ . A larger P allows greater initial uncertainty ϵ and the resulting closed loop system has favorable time domain performance [22]–[13].

However, for the time delay systems it should be noted that the effect of initial uncertainties in the input delay segment $u_1(\cdot)$ is independent to the choice of P and causes definite damage to the time domain performance.

The operator Riccati equation (12) has similar structure to those arise in LQ problem for delay systems [5]. In [5] it is shown that the solution $S > 0$ ($S \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$) has a form of an integral kernel representation $\{S_0, S_1, S\}$

$$\begin{aligned} (S\alpha)^0 &= S_0\alpha + \int_{-h}^0 S_1(t)\alpha^1(t)dt \\ (S\alpha)^1(\alpha) &= S_1^T(\alpha)\alpha^0 + \int_{-h}^0 S_2(\alpha, t)\alpha^1(t)dt \\ &-h \leq \alpha \leq 0 \end{aligned}$$

With the integral kernel representation of operator $S \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$ the control law (13) is described by

$$\begin{aligned} u(t) &= K_0 \underline{x}(t) + \int_{-h}^0 K_1(t) u(t+\tau) d\tau \\ \underline{x}(t) &= A \underline{x}(t) + B u(t-h) + \mathcal{P}\mathcal{C}^T(u(t) - \mathcal{C} \underline{x}(t)) \\ K_0 &= -S_1^T(0), \quad K_1(\cdot) = -S_2(\cdot, 0) \quad (21) \end{aligned}$$

and has a feature of observer-based predictive controller. The feedback gains K_0 and $K_1(\cdot)$ are the parameters defined by the integral kernel representation $\{S_0, S_1, S_2\}$.

Next we will show that the stabilizing solution $S > 0$ ($S \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$) is constructively given based on finite-dimensional Riccati equation. The following result enables us to avoid the direct calculation of (12) (see also [23]).

Theorem 4: Let $\mathcal{M} > 0$ ($\mathcal{M} \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$) be the stabilizing solution to the operator Riccati equation

$$\mathcal{M}A\phi + A^*\mathcal{M}\phi - \mathcal{M}BB^*\mathcal{M}\phi + C^*C\phi = 0, \quad \phi \in \mathcal{X} \quad (22)$$

then, for given constant $\gamma > (1 + \lambda_{\max}(\mathcal{M}\mathcal{P}))^{1/2}$, the stabilizing solution $S > 0$ to (12) is described by

$$S := \left(I - \frac{1}{\gamma^2} \cdot (I + \mathcal{M}\mathcal{P}) \right)^{-1} \mathcal{M}.$$

Furthermore, based on a solution $M > 0$ to the finite-dimensional Riccati equation

$$MA + A^T M - MBB^T M + C^T C = 0 \quad (23)$$

the stabilizing solution $\mathcal{M} > 0$ ($\mathcal{M} \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$) to (22) is constructively given as follows

$$(\mathcal{M}\phi)^0 := M_0\phi^0 + \int_{-h}^0 M_1(\beta)\phi^1(\beta) d\beta \quad (24)$$

$$(\mathcal{M}\phi)^1(\alpha) := M_1^T(\alpha)\phi^0 + \int_{-h}^0 M_2(\alpha, \beta)\phi^1(\beta) d\beta, \quad -h \leq \alpha \leq 0 \quad (25)$$

$$M_0 = e^{A^T h} M e^{A h} + \int_{-h}^0 e^{A^T(\xi+h)} C^T C e^{A(\xi+h)} d\xi$$

$$M_1(\beta) = e^{A^T h} M e^{-A\beta} B + \int_{\beta}^0 e^{A^T(\xi+h)} C^T C e^{A(\xi-\beta)} d\xi$$

$$M_2(\alpha, \beta) = B^T e^{-A^T \alpha} M e^{-A\beta} B + \int_{\max(\alpha, \beta)}^0 B^T e^{A^T(\xi-\alpha)} C^T C e^{A(\xi-\beta)} B d\xi.$$

Proof: Based on (11) and (22), it is verified that the equality

$$\begin{aligned} & \mathcal{M} \left(A + \frac{1}{\gamma^2 - 1} \cdot PC^*C \right) \left(I - \frac{1}{\gamma^2} \cdot (I + \mathcal{P}\mathcal{M}) \right) \phi \\ & + \left(I - \frac{1}{\gamma^2} \cdot (I + \mathcal{M}\mathcal{P}) \right) \left(A + \frac{1}{\gamma^2 - 1} \cdot PC^*C \right)^* \mathcal{M}\phi \\ & - \mathcal{M}BB^*\mathcal{M}\phi + \frac{1}{\gamma^2 - 1} \cdot \mathcal{M}PC^*CP\mathcal{M}\phi + \frac{\gamma^2}{\gamma^2 - 1} \\ & \cdot \left(I - \frac{1}{\gamma^2} \cdot (I + \mathcal{M}\mathcal{P}) \right) C^*C \left(I - \frac{1}{\gamma^2} \cdot (I + \mathcal{P}\mathcal{M}) \right) \\ & \cdot \phi = 0 \end{aligned}$$

holds for $\phi \in \mathcal{X}$. Hence, interpreting $\mathcal{P}\mathcal{M}$ and $\mathcal{M}\mathcal{P}$ on \mathcal{X} , the operator $S := (I - (1/\gamma^2) \cdot (I + \mathcal{M}\mathcal{P}))^{-1} \mathcal{M}$ with the stabilizing solutions $\mathcal{M} > 0$, $\mathcal{P} > 0$ is shown to be positive definite for given $\gamma > (1 + \lambda_{\max}(\mathcal{M}\mathcal{P}))^{1/2}$. It follows from the equality

$$\begin{aligned} & \left(A - BB^* + \frac{1}{\gamma^2 - 1} \cdot PC^*C(I + \mathcal{P}S) \right) \\ & \cdot \left(I - \frac{1}{\gamma^2} \cdot (I + \mathcal{P}\mathcal{M}) \right) \phi \\ & = \left(I - \frac{1}{\gamma^2} \cdot (I + \mathcal{P}\mathcal{M}) \right) (A - B^*B\mathcal{M})\phi \end{aligned}$$

that the operator $S > 0$ is a stabilizing solution.

Now, we will verify that the parameters $\{M_0, M_1, M_2\}$ meet the integral kernel representation of operator \mathcal{M} . Using the representations (24) and (25), (22) is written as partial differential equations

$$M_0 A + A^T M_0 - M_1(0)M_1^T(0) + C^T C = 0, \quad M_0 = M_0^T \quad (26)$$

$$-\frac{d}{d\beta} M_1(\beta) + A^T M_1(\beta) - M_1(0)M_2(0, \beta) = 0, \quad (27)$$

$$-\left(\frac{\partial}{\partial \alpha} + \frac{\partial}{\partial \beta} \right) M_2(\alpha, \beta) - M_2(\alpha, 0)M_2(0, \beta) = 0, \quad M_2(\alpha, \beta) = M_2^T(\beta, \alpha). \quad (28)$$

And further, since $\mathcal{M}\phi \in \mathcal{V}^* = \mathcal{D}(A^*)$, the following boundary conditions are required

$$M_1(-h) = M_0 B, \quad M_2(-h, \beta) = B^T M_1(\beta) \quad (-h \leq \beta \leq 0). \quad (29)$$

Hence, substituting directly, it is verified that the parameters $\{M_0, M_1, M_2\}$ meet the integral kernel representation of operator $\mathcal{M} \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$.

Next, we will show that the solution $\{M_0, M_1, M_2\}$ is positive definite and the stabilizing solution. Since the inequality

$$\begin{aligned} \langle \phi, \mathcal{M}\phi \rangle &= m^T(0) M m(0) + \int_{-h}^0 m^T(\xi) C^T C m(\xi) d\xi > 0 \\ m(\xi) &:= e^{A(\xi+h)} \phi^0 + \int_{-h}^{\xi} e^{A(\xi-\beta)} B \phi^1(\beta) d\beta \end{aligned}$$

holds for all $\phi \neq 0$ ($\phi \in \mathcal{V}$) (note that (C, A, B) is observable and controllable), the operator \mathcal{M} is shown to be positive definite. While the system

$$\begin{aligned} \dot{p}(t) &= Ap(t) + Bv(t-h) \\ v(t) &= -B^T M \left\{ e^{A h} p(t) + \int_{-h}^0 e^{-A\beta} B v(t+\beta) d\beta \right\} \end{aligned}$$

associated with the infinitesimal generator $A - BB^*M$ has stable poles $\lambda(A - BB^*M)$, hence the operator \mathcal{M} is shown to be the stabilizing solution to (22). Q.E.D.

Remark 4: The value $(1 + \lambda_{\max}(\mathcal{M}\mathcal{P}))^{1/2}$ stated in Theorem 4 qualifies the alternative analytic representation of the maximum robustness margin ϵ_{opt} (see [3]). Since the operator $\mathcal{M}\mathcal{P}$ is compact, it is also possible to approach the value ϵ_{opt} based on the finite-rank approximation of operator $\mathcal{M}\mathcal{P}$.

By Theorem 4, it is shown that the operator S is constructively given based on finite-dimensional Riccati equations. However it should be noted that, in return for avoiding the direct calculation of (12), it arises infinite-dimensional manipulation concerning with the compact operator $\mathcal{M}\mathcal{P}$. Based on the previous results stated in Theorems 2 and 4, the design procedure is summarized in the following way.

Theorem 5 (Main Result): For a prescribed margin of the required robust stability ϵ ($\epsilon < \epsilon_{opt} < 1$, see Fact 1), the robust controller

$$\begin{aligned} u(t) &= K_0 \underline{x}(t) + \int_{-h}^0 K_1(\beta) u(t+\beta) d\beta \\ \dot{\underline{x}}(t) &= A \underline{x}(t) + Bu(t-h) + L(y(t) - C \underline{x}(t)) \end{aligned}$$

is obtained based on the following procedure.

Step 1: Calculate the solutions $M > 0$, $P > 0$ to the finite-dimensional Riccati equations

$$\begin{aligned} MA + A^T M - MBB^T M + C^T C &= 0, \\ PA^T + AP - PC^T CP + BB^T &= 0 \end{aligned}$$

and let $L := PC^T$.

Step 2 Define the operators $\mathcal{M} \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$ (Theorem 4) and $\mathcal{P} \in \mathcal{L}(\mathcal{V})$ (Lemma 3) calculate the following operator

$$\begin{aligned} S &= (I - \epsilon^2 (I + \mathcal{M}\mathcal{P}))^{-1} \mathcal{M} \\ &= \frac{1}{1 - \epsilon^2} (I - T)^{-1} \mathcal{M} \\ T &= \frac{\epsilon^2}{1 - \epsilon^2} \mathcal{M}\mathcal{P} \end{aligned} \quad (30)$$

Since the operator $T \in \mathcal{L}(\mathcal{V})$ is compact, we can construct a sequence of finite-rank operators $\{T_n\}$ which converge to T in the sense of the uniform operator topology $\|T - T_n\| \rightarrow 0$. It follows from the fact $\|(I - T_n)^{-1} - (I - T)^{-1}\| \rightarrow 0$ that the sequence $\{S_n\}$, $S_n = (1/(1 - \epsilon^2)) (I - T_n)^{-1} \mathcal{M}$ also converges to S in the sense of the uniform operator topology.

On the finite rank approximation of operator T the averaging method [1], [9] is applicable. In this case the manipulation of operator $(I - T)^{-1}$ is reduced to the calculation of inverse matrix.

Step 3 Based on the integral kernel representation of $S \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$ determine the feedback gains $K_0, K_1(\cdot)$

$$K = -S_1^*(0), \quad K_1(\cdot) = -S_2(0, \cdot)$$

In case we approximate the operator S by averaging method, say S_n , the parameters $S_1^*(0), S_2(0, \cdot)$ are determined by partitioning the corresponding matrix form representation of S .

IV. CONCLUSION

A robust stabilization problem in the gap metric is discussed for a system with delays in control. We derive a design procedure of robust controllers based on finite-dimensional Riccati equations and it is shown that the resulting controller has a structure of observer based predictive actions. The completing the square argument employed in the derivation enables us to provide game theoretic interpretations on the trade off between the initial uncertainties and attenuating the disturbance.

The approach discussed here is applicable for the H^∞ output feedback problem in more general setting and it is also shown that the admissible control law is constructively given in the framework of finite dimensional operations.

REFERENCES

- [1] H. T. Banks and J. A. Burns, 'Hereditary control problems: Numerical methods based on averaging approximations', *SIAM J. Contr. Optim.* vol. 16, no. 2, pp. 164-208, 1978.
- [2] T. Başar and P. Bernhard, *H^∞ Optimal Control And Related Minimax Design Problems: A Dynamic Game Approach*, Boston: Birkhauser, 1991.
- [3] R. I. Curtain, 'Robust stabilizability of normalized coprime factors: The infinite dimensional case', *Int. J. Contr.* vol. 51, no. 6, pp. 1173-1190, 1990.
- [4] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory* (Lecture Notes in Control and Information Sciences) vol. 8, Berlin: Springer Verlag, 1978.
- [5] M. C. Delfour, 'The linear quadratic optimal control problem with delays in state and control variables: A state space approach', *SIAM J. Contr. Optim.* vol. 24, no. 5, pp. 121-144, 1986.
- [6] H. Dym, T. Georgiou and M. C. Smith, 'Direct design of optimal controllers for delay systems', in *Proc. 32nd IEEE CDC*, 1993, pp. 3821-3822.
- [7] T. Georgiou and M. C. Smith, 'Optimal robustness in the gap metric', *IEEE Trans. Automat. Contr.* vol. 35, pp. 673-686, 1990.
- [8] —, 'Robust stabilization in the gap metric: Controller design for distributed plants', *IEEE Trans. Automat. Contr.* vol. 37, no. 8, pp. 1133-1143, 1992.
- [9] J. S. Gibson, 'Linear quadratic optimal control of hereditary differential systems: Infinite dimensional Riccati equations and numerical approximations', *SIAM J. Contr. Optim.* vol. 21, no. 1, pp. 95-139, 1983.
- [10] A. Ichikawa, 'Quadratic control of evolution equations with delays in control', *SIAM J. Contr. Optim.* vol. 20, no. 5, pp. 645-668, 1982.
- [11] —, H^∞ control and min max problems in Hilbert space, Dept. Electrical Eng., Shizuoka University, Japan Tech. Rep. C-17, 1991.
- [12] T. Kato, *Perturbation Theory for Linear Operators*, New York: Springer Verlag, 1966.
- [13] A. Kojima, M. Fujita, K. Uchida and E. Shimemura, 'Linear quadratic differential game and H^∞ control—A direct approach based on completing the square', *SICE Trans.* (in Japanese) vol. 28, no. 5, pp. 570-577, 1992.
- [14] A. Kojima and S. Ishijima, 'Robust stabilization problem for time delay systems based on spectral decomposition approach', in *Proc. 1992 Amer. Contr. Conf.*, Chicago, IL, 1992, pp. 998-1003.
- [15] —, 'Robust stabilization problem for a system with delays in control—A correspondence of controller design case study for ill conditioned distillation system', in *Proc. 12th IFAC World Congress*, vol. 2, Sydney, Australia, 1993, pp. 279-283.
- [16] A. Kojima, K. Uchida and E. Shimemura, 'Robust stabilization of uncertain time delay systems via combined internal external approaches', *IEEE Trans. Automat. Contr.* vol. 38, no. 2, pp. 373-378, 1993.
- [17] D. C. McFarlane and K. Glover, *Robust Controller Design Using Normalized Coprime Factor Plant Descriptions* (Lecture Notes in Control and Information Sciences) vol. 138, Berlin: Springer Verlag, 1990.
- [18] —, 'A loop shaping design procedure using H^∞ synthesis', *IEEE Trans. Automat. Contr.* vol. 37, no. 6, pp. 759-769, 1992.
- [19] J. R. Partington and K. Glover, 'Robust stabilization of delay systems by approximation of coprime factors', *Syst. Contr. Lett.* vol. 14, pp. 325-331, 1990.
- [20] A. I. Pritchard and D. Salamon, 'The linear quadratic control problem for retarded systems with delays in control and optimization', *IMA J. Math. Contr. Inform.* vol. 2, pp. 335-361, 1985.
- [21] —, 'The linear quadratic control problem for infinite dimensional systems with unbounded input and output operators', *SIAM J. Contr. and Optim.* vol. 25, no. 1, pp. 121-144, 1987.
- [22] K. Uchida and M. Fujita, 'On the central controller: Characterizations via differential games and LQG control problem', *Syst. Contr. Lett.* vol. 13, pp. 9-13, 1989.
- [23] K. Uchida and E. Shimemura, 'Closed loop properties of the infinite time linear quadratic optimal regulator for systems with delays', *Int. J. Contr.* vol. 43, no. 3, pp. 773-779, 1986.
- [24] B. van Keulen, *H^∞ control For Distributed Parameter Systems: A State Space Approach*, Boston: Birkhauser, 1993.

A New Balanced Canonical Form for Stable Multivariable Systems

Bernard Hanzon

1. INTRODUCTION

A new balanced canonical form is presented for stable multivariable linear systems. In [3] overlapping continuous block balanced canonical forms were introduced for the stable single input/single output (SISO) case as a generalization of the balanced canonical form of [9] for the SISO case. In the search for a generalization of these results to the multivariable case a new multivariable balanced

Manuscript received March 15, 1993; revised January 4, 1994.

The author is with the Department of Econometrics, Free University, Amsterdam, Holland.

IEEE Log Number 9407356.

canonical form was discovered, which is of interest in its own right and which will be presented here. The new canonical form has a number of nice properties. The integer invariants that appear in the canonical form are the multiplicities of the Hankel singular values and a number of new invariants, which are in one-to-one bijective correspondence with the Kronecker indexes of subsystems. Truncation of the state vector leads to stable minimal models in canonical form, just as in the case of [9]. In the SISO case the canonical form coincides with Ober's balanced canonical form (up to a different signs convention). The reachability matrix of a system in canonical form with identical singular values is positive upper triangular. For the subclass of stable multivariable all-pass systems a detailed treatment of the canonical form is presented in [4], Section IV.

II. BALANCING, CANONICAL FORMS, AND KRONECKER INDEXES

Let us consider continuous-time multivariable systems of the form

$$\dot{x}_t = Ax_t + Bu_t, \quad (2.1)$$

$$y_t = Cx_t + Du_t \quad (2.2)$$

with $t \in \mathbb{R}$, $u_t \in \mathbb{R}^p$, $x_t \in \mathbb{R}^n$, $y_t \in \mathbb{R}^m$, $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times p}$, $C \in \mathbb{R}^{m \times n}$, $D \in \mathbb{R}^{m \times p}$.

Let for each $n \in \{1, 2, 3, \dots\}$ the set C_n be the set of all quadruples $(A, B, C, D) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times p} \times \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times p}$ with the properties: a) (A, B, C, D) is a minimal realization and b) the spectrum of A is contained in the open left-half plane (i.e., A is asymptotically stable). A quadruple (A, B, C, D) or triple (A, B, C) will be called asymptotically stable iff the spectrum of A lies in the open left-half plane (irrespective of minimality or nonminimality).

As is well known, two minimal system representations (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) have the same transfer function $G(s) = C_1(sI - A_1)^{-1}B_1 + D_1 = C_2(sI - A_2)^{-1}B_2 + D_2$, and therefore describe the same input-output behavior, iff there exists an $n \times n$ matrix $T \in GL_n(\mathbb{R})$ such that $A_1 = TA_2T^{-1}$, $B_1 = TB_2$, $C_1 = C_2T^{-1}$, $D_1 = D_2$. In that case we say that (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) are i/o-equivalent. This is clearly an equivalence relation; write $(A_1, B_1, C_1, D_1) \sim (A_2, B_2, C_2, D_2)$. A unique representation of a linear system can be obtained by deriving a canonical form.

Definition 2.1: A canonical form for an equivalence relation " \sim " on a set X is a map

$$\Gamma: X \rightarrow X$$

which satisfies for all $x, y \in X$:

- i) $\Gamma(x) \sim x$
- ii) $x \sim y \iff \Gamma(x) = \Gamma(y)$

Equivalently a canonical form can be given by the image set $\Gamma(X)$; a subset $B \subseteq X$ describes a canonical form if for each $x \in X$ there is precisely one element $b \in B$ such that $b \sim x$. The mapping $x \mapsto b \in B$, $x \mapsto b$ then describes a canonical form.

Let $(A, B, C, D) \in C_n$. The controllability Grammian W_c is the positive definite symmetric matrix that is given by the integral

$$W_c = \int_0^\infty \exp(At)BB^T \exp(A^T t) dt.$$

As is well known, W_c can be obtained as the unique solution of the following Lyapunov equation

$$AW_c + W_cA^T = -BB^T. \quad (2.3)$$

On a dual fashion, the observability Grammian W_o is the positive definite symmetric matrix that is given by the integral

$$W_o = \int_0^\infty \exp(A^T t)C^T C \exp(At) dt.$$

This matrix is the unique solution of the following Lyapunov equation

$$A^T W_o + W_o A = -C^T C. \quad (2.4)$$

Definition 2.2: Let $(A, B, C, D) \in C_n$, then (A, B, C, D) is called balanced if the corresponding observability and controllability Grammians are equal and diagonal, i.e., there exist positive numbers $\sigma_1, \sigma_2, \dots, \sigma_n$ such that

$$W_o = W_c = \text{diag}(\sigma_1, \dots, \sigma_n) =: \Sigma. \quad (2.5)$$

The numbers $\sigma_1, \dots, \sigma_n$ are called the (Hankel) singular values of the system. It will be convenient to call an arbitrary quadruple $(A, B, C, D) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times p} \times \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times p}$ balanced if the pair of Lyapunov equations $A\Sigma + \Sigma A^T = -BB^T$, $A^T \Sigma + \Sigma A = -C^T C$ has a positive definite solution of the form $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ (assuming neither asymptotic stability nor minimality).

By a suitable permutation of the components of the state of a balanced system, a representation of the system is obtained for which the singular values are ordered with respect to size: $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n > 0$. The singular values, ordered in this way, are known to be uniquely determined by the input-output behaviour of the system.

Definition 2.3: A balanced canonical form (on C_n) is a canonical form $\Gamma: C_n \rightarrow C_n$, such that $\Gamma(A, B, C, D)$ is balanced for each quadruple $(A, B, C, D) \in C_n$.

Definition 2.4: Consider a pair (A, B) of matrices $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times p}$. Let $R_n = R_n(A, B) = [B, AB, \dots, A^{n-1}B]$ denote the corresponding reachability matrix. Suppose the pair (A, B) is reachable, i.e., the reachability matrix has rank n . The selection of the first n linearly independent columns is called the Kronecker selection. It has the property that it is a so-called nice selection, which means that if the j th column of R_n is in the selection, then either $j \leq p$ or otherwise the $(j-p)$ th column is also in the selection. For each $i \in \{1, 2, \dots, p\}$ let d_i denote the smallest nonnegative value of j such that the $(jp+i)$ th column is not in the selection. Then we will call (d_1, d_2, \dots, d_p) the dynamical indexes (also called successor indexes) corresponding to the selection. By ordering these according to magnitude, one obtains a nondecreasing sequence of p indexes $\kappa_1 \leq \kappa_2 \leq \dots \leq \kappa_p$ which are called the Kronecker reachability (or controllability) indexes.

With any nice selection corresponds a sequence of integers $p = s_0 \geq s_1 \geq s_2 \geq \dots \geq s_l > s_{l+1} = 0$ which add up to $n+p$ and a sequence of sets of indexes $\{\{i_j(1), i_j(2), \dots, i_j(s_j)\} \subset \{1, 2, \dots, s_{j-1}\}, j = 1, 2, \dots, l\}$ with the property that of the s_{j-1} columns that can be chosen from $A^{j-1}B$ in the nice selection the $i_j(1)$ -th, the $i_j(2)$ -th etc until the $i_j(s_j)$ -th are chosen. Because the Kronecker selection is also a nice selection these quantities are also defined for the Kronecker selection. It is clear that the sequence of sets of indexes determines the Kronecker selection completely and is in bijective correspondence with the sequence of dynamical indexes $\{d_1, d_2, \dots, d_p\}$ that describes the Kronecker selection. It is well-known and can easily be derived from the foregoing that the Kronecker indexes are in one-to-one bijective correspondence with the sequence $\{s_n\}_{n=1}^l$ (cf. e.g., [2]).

Remark: A similar definition holds for the Kronecker selection of rows from the observability matrix of a pair of matrices $(A, C) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{m \times n}$, the corresponding Kronecker observability indexes etc.

The following lemma is basic for our considerations (see e.g., [9]):

Lemma 2.5. Let $M \in \mathbb{R}^{n \times l}$, $\text{rank}(M) = n \leq l$. There exists an orthogonal matrix $Q_0 \in \mathbb{R}^{n \times n}$ and natural numbers $1 \leq$

$i_1 < i_2 < \dots < i_n \leq l$ such that

$$M_0 := Q_0 M = \begin{pmatrix} 0 & \dots & m_{1i_1} & * & \dots & * & \dots & \dots & * \\ 0 & \dots & 0 & 0 & \dots & m_{2i_2} & * & \dots & * \\ \vdots & & \vdots & \vdots & & \vdots & & & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & m_{ni_n} & * \end{pmatrix}$$

with $m_{ji} = 0$ for all $i < i_j$ and $m_{ji} > 0$ for all $j \in 1, 2, \dots, n$; M_0 is unique and Q_0 is unique. Such a matrix will be called positive upper triangular with independency indexes i_1, i_2, \dots, i_n .

A matrix will be called full rank upper triangular if it is positive upper triangular up to multiplication of some (or possibly all or none) of its rows by -1.

III. A BALANCED CANONICAL FORM FOR SYSTEMS WITH IDENTICAL SINGULAR VALUES

The following theorem of [1] is basic for the relation between systems with all singular values equal and stable all-pass systems.

Theorem 3.1: Let $m = p$.

- If a balanced triple (A, B, C) has identical Hankel singular values $\sigma = 1$ then there exists an orthogonal matrix D such that $C = -DB^T$.
- (A, B, C, D) is a balanced asymptotically stable realization of a stable all-pass system iff (A, B, C) is a balanced asymptotically stable triple with identical singular values and D is an orthogonal matrix such that $C = -DB^T$.

It will be useful to extend the usual definition of orthogonal square matrices to rectangular matrices.

Definition 3.2: An $m \times p$ matrix U will be called orthogonal if $U^T U = I_p$ (if $p \leq m$) or $U U^T = I_m$ (if $m \leq p$).

Using this, a balanced realization (A, B, C) of a stable system with identical singular values (with possibly $m \neq p$) can be characterized as follows:

Corollary 3.3: The following three statements are equivalent:

- A triple (A, B, C) is a balanced asymptotically stable realization of a stable system with identical singular values $\sigma > 0$.
- The pair (A, B) is reachable and $A + A^T = -(1/\sigma)BB^T = -(1/\sigma)C^T C$, $\sigma > 0$.
- The pair (A, B) is reachable, $A + A^T = -(1/\sigma)BB^T$, and there exists a (possibly rectangular) orthogonal matrix D such that $C = -DB^T$.

Proof: Without loss of generality one can assume that $m = p$, because if $m \neq p$ than one can add a sufficient number of zero rows to C or zero columns to B to obtain a system with the same number of in- and outputs and clearly if the result holds for the square system obtained in this way, it also holds for the original system. If $m = p$ the theorem follows from the previous theorem together with a theorem of [10], (here applied to the special case where the Lyapunov equation involved has the identity matrix as a solution) which says that if (A, B) satisfies the equation $A + A^T = -BB^T$ then: A is asymptotically stable iff (A, B) is reachable. \square

The following canonical form for the set of stable multivariable all-pass systems of fixed McMillan degree will be important for our constructions.

Theorem 3.4 ([4], Section IV): The following two statements are equivalent: i) A system Π is a stable all-pass system with McMillan degree n . ii) There exists a unique balanced realization $(A, B, C, D) \in \mathcal{C}_n$ of Π of the following form: There are integers $p = s_0 \geq s_1 \geq s_2 \geq \dots \geq s_l > s_{l+1} = 0$ which add up to $n + p$, such that

$$B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}$$

where B_1 is an $s_1 \times s_0$ positive upper triangular matrix;

$$A = \begin{pmatrix} A_{11} & A_{12} & 0 & \dots & 0 \\ A_{21} & A_{22} & A_{23} & \ddots & \vdots \\ 0 & A_{32} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & A_{l-1,l} \\ 0 & \dots & 0 & A_{l,l-1} & A_{l,l} \end{pmatrix}$$

a block tridiagonal matrix with $A_{u,v}$ an $s_u \times s_v$ matrix, $u, v \in \{1, 2, \dots, l\}$, $A_{u,v} = 0$ if $|u - v| > 1$;

$$A_{11} = \tilde{A}_{11} - \frac{1}{2} B_1 B_1^T$$

\tilde{A}_{11} an otherwise arbitrary skew symmetric $s_1 \times s_1$ matrix; $A_{uu} = \tilde{A}_{uu}$ an otherwise arbitrary skew symmetric $s_u \times s_u$ matrix for each $u \in \{2, 3, \dots, l\}$; $A_{u+1,u}$ a positive upper triangular $s_{u+1} \times s_u$ matrix for each $u \in \{1, 2, \dots, l-1\}$; $A_{u,u+1} = -A_{u+1,u}^T$ for each $u \in \{1, 2, \dots, l-1\}$; D an otherwise arbitrary orthogonal $p \times p$ matrix and

$$C = -DB^T.$$

The indexes $s_u, u = 1, \dots, l$ are in bijective correspondence with the Kronecker indexes. The canonical form is balanced and its reachability matrix is positively upper triangular.

For triples (A, B, C) with all Hankel singular values equal to unity, possibly with $m \neq p$, one obtains the same balanced canonical form with the only exception that, while in the case of all-pass systems the matrix C can be determined from B and D , here instead C is an arbitrary solution of the equation

$$C^T C = B B^T = \begin{pmatrix} B_1 B_1^T & 0 \\ 0 & 0 \end{pmatrix}$$

(Note that because $B_1 B_1^T$ is an $s_1 \times s_1$ positive definite matrix, one has $s_1 \leq m, s_1 \leq p$.) It follows that the matrix C can be partitioned as $[C_1, 0]$, C_1 an $m \times s_1$ matrix which is a solution of $C_1^T C_1 = B_1 B_1^T$. Let $\tilde{B}_1 = [(B_1 B_1^T)^{1/2}, 0]$, where the zero matrix is $s_1 \times (p - s_1)$ and $(B_1 B_1^T)^{1/2}$ is the $s_1 \times s_1$ positive definite symmetric square root of $B_1 B_1^T$. Then clearly $C_1^T C_1 = \tilde{B}_1 \tilde{B}_1^T$ and $C^T C = \tilde{B} \tilde{B}^T$ with $\tilde{B} = \begin{pmatrix} \tilde{B}_1 \\ 0 \end{pmatrix}$ an $m \times s_1$ matrix. From the corollary above it follows that there exists an $m \times s_1$ orthogonal matrix D such that $C = -D \tilde{B}^T$ and so $C_1 = -D \tilde{B}_1^T$. Partition $D = [-U, V]$ where U is an $m \times s_1$ orthogonal matrix and V is an $m \times (p - s_1)$ orthogonal matrix. It follows that $C_1 = U (B_1 B_1^T)^{1/2}$. Because $(B_1 B_1^T)^{1/2}$ is positive definite, the relation between C_1 and U is bijective: $U = C_1 (B_1 B_1^T)^{-1/2}$ and therefore U can be used to parametrize C_1 . In this way one obtains a parametrization of the canonical form for systems with one unit singular value: The matrices B and A together have $s_1 + s_2 + \dots + s_l = n$ positive parameters due to the requirement that $B_1, A_{21}, A_{32}, \dots, A_{l,l-1}$ are all positive upper triangular, while all other entries in these positive upper triangular matrices are either prescribed to be zero by the structural indexes, or are free to vary over the real numbers; furthermore the skew symmetric matrices $\tilde{A}_{11}, \tilde{A}_{22}, \dots, \tilde{A}_{l,l}$ have $\sum_{u=1}^l (1/2) s_u (s_u - 1)$ parameters that are free to vary over the real numbers and finally $C_1 = U (B_1 B_1^T)^{1/2}$ is parametrized by the $m \times s_1$ orthogonal matrix U ; the set of all such orthogonal matrices has dimension $s_1 m - (1/2) s_1 (s_1 + 1)$. Of course one could add a feedthrough matrix to the system, which would have mp freely varying real parameters.

If (A, B, C) has all singular values equal to $\sigma > 0$, then $(\bar{A}, \bar{B}, \bar{C}) = (A, B/\sqrt{\sigma}, C/\sqrt{\sigma})$ has all its singular values equal to unity. Requiring $(\bar{A}, \bar{B}, \bar{C})$ to be in the canonical form just described leads to a canonical form for (A, B, C) by taking $(A, B, C) =$

$(\bar{A}, \sqrt{\sigma} \bar{B}, \sqrt{\sigma} \bar{C})$. This canonical form will be used in the sequel. Clearly $\bar{B} \bar{B}^T = B B^T / \sigma$, $\bar{C} \bar{C}^T = C C^T / \sigma$, etc.

IV. A NEW BALANCED CANONICAL FORM FOR STABLE MULTIVARIABLE SYSTEMS

Consider a system with balanced state space representation (A, B, C, D) . The Grammians are equal and of the form

$$\Sigma = \text{diag}(\sigma_1 I_{n(1)}, \sigma_2 I_{n(2)}, \dots, \sigma_k I_{n(k)})$$

where $\sigma_1 > \sigma_2 > \dots > \sigma_k > 0$ are the Hankel singular values of the system and $n(1), n(2), \dots, n(k)$ the corresponding multiplicities. Partition A, B, C , according to $n(1), n(2), \dots, n(k)$ to obtain

$$A = \begin{pmatrix} A(1,1) & A(1,2) & \dots & A(1,k) \\ A(2,1) & A(2,2) & \dots & A(2,k) \\ \vdots & \vdots & \ddots & \vdots \\ A(k,1) & A(k,2) & \dots & A(k,k) \end{pmatrix} \quad (4.6)$$

and

$$B = \begin{bmatrix} B(1) \\ \vdots \\ B(k) \end{bmatrix}, \quad C = [C(1) \ C(2) \ \dots \ C(k)]. \quad (4.7)$$

The following result is very important for the construction of balanced canonical forms.

Theorem 4.1 ([10], [7]): Let (A, B, C, D) be partitioned as above, balanced in the sense of definition 2.2, but not necessarily asymptotically stable and minimal. Then $(A, B, C, D) \in C_n$ iff for each $i \in \{1, 2, \dots, k\}$, $(A(i, i), B(i), C(i), D) \in C_{n(i)}$.

The balancing (Lyapunov) equations for (A, B, C, D) are given in terms of the $A(i, j), B(i), C(j), i = 1, \dots, k, j = 1, \dots, k$ by

$$\begin{aligned} A(i, i) + A(i, i)^T &= -\frac{1}{\sigma_i} B(i) B(i)^T \\ &= -\frac{1}{\sigma_i} C(i)^T C(i) \end{aligned} \quad (4.8)$$

$$A(i, j) \sigma_j + A(j, i)^T \sigma_i = -B(i) B(j)^T \quad (4.9)$$

$$A(i, j) \sigma_i + A(j, i)^T \sigma_j = -C(i)^T C(j) \quad (4.10)$$

where $i \in \{1, \dots, k\}, j \in \{1, \dots, k\}, i \neq j$. Note that the last two equations can be solved in terms of $A(i, j)$ and $A(j, i)$ for given pair $(i, j), i \neq j$ and given $B(i), B(j), C(i), C(j)$, because $\sigma_i \neq \sigma_j$. Therefore, and because of the theorem 4.1, the construction of a balanced canonical form can be reduced to constructing a balanced canonical form for the subsystems $(A(i, i), B(i), C(i)), i = 1, 2, \dots, k$. For each $i \in \{1, \dots, k\}$ such a system is a system with identical singular values, or perhaps one should say with one singular value σ_i . For such systems the canonical form presented in Section III can be used. This leads to the following result.

Theorem 4.2:

a) The following statements are equivalent:

- i) Ξ is a stable input-output-system with p inputs and m outputs and McMillan degree n .

- ii) Ξ has a realization $(A, B, C, D) \in C_n$ of the following form: There are numbers $\sigma_1 > \sigma_2 > \dots > \sigma_k > 0$, the Hankel singular values, and integers $n(1), n(2), \dots, n(k)$, the corresponding multiplicities of the singular values, and for each $i \in \{1, 2, \dots, k\}$ there are integers $p = s_0(i) \geq s_1(i) \geq \dots \geq s_{l(i)}(i) > s_{l(i)+1}(i) = 0$ which add up to $n(i) + p$, such that

$$A = \begin{pmatrix} A(1,1) & A(1,2) & \dots & A(1,k) \\ A(2,1) & A(2,2) & \dots & A(2,k) \\ \vdots & \vdots & \ddots & \vdots \\ A(k,1) & A(k,2) & \dots & A(k,k) \end{pmatrix} \quad (4.11)$$

and

$$B = \begin{bmatrix} B(1) \\ \vdots \\ B(k) \end{bmatrix}, \quad C = [C(1) \ C(2) \ \dots \ C(k)] \quad (4.12)$$

where $A(i, j)$ is an $n(i) \times n(j)$ matrix, $B(i)$ an $n(i) \times p$ matrix and $C(j)$ an $m \times n(j)$ matrix, $i = 1, 2, \dots, k, j = 1, 2, \dots, k$ and the triples $(A(i, i), B(i), C(i))$ have the following form

$$B(i) = \begin{bmatrix} B_1(i) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (4.13)$$

where $B_1(i)$ is an $s_1(i) \times p$ positive upper triangular matrix and see also (4.14), found at the bottom of the page, a block tridiagonal matrix where for each $u, v \in \{1, 2, \dots, l(i)\}$, $A_{u,v}(i, i)$ is an $s_u(i) \times s_v(i)$ matrix, $A_{u,v}(i, i) = 0$ if $|u - v| > 1$; $A_{11}(i, i) = \tilde{A}_{11}(i, i) - (1/2) B_1(i) B_1(i)^T / \sigma_i$, where $\tilde{A}_{11}(i, i)$ is an otherwise arbitrary skew symmetric $s_1(i) \times s_1(i)$ matrix; $A_{uu}(i, i) = \tilde{A}_{uu}(i, i)$ an otherwise arbitrary skew symmetric $s_u(i) \times s_u(i)$ matrix for each $u \in \{2, 3, \dots, l(i)\}$; $A_{u+1,u}(i, i)$ a positive upper triangular $s_{u+1}(i) \times s_u(i)$ matrix for each $u \in \{1, 2, \dots, l(i) - 1\}$; $A_{u,u+1}(i, i) = -A_{u+1,u}(i, i)^T$ for each $u \in \{1, 2, \dots, l(i) - 1\}$, and furthermore

$$C(i) = [C_1(i), 0, \dots, 0] \quad (4.15)$$

where

$$C_1(i) = U(i) (B_1(i) B_1(i)^T)^{1/2} \quad (4.16)$$

in which $U(i)$ is an $m \times s_1(i)$ orthogonal matrix, i.e., $U(i)^T U(i) = I_{s_1(i)}$; furthermore the matrices $A(i, j), i \neq j, j \in \{1, \dots, k\}$ are determined as the unique solution of the linear equations (4.9), (4.10).

- b) For each system Ξ the the realization $(A, B, C, D) \in C_n$ described in ii) above is unique and balanced. This determines a balanced canonical form for i/o-equivalence on C_n .

$$A(i, i) = \begin{pmatrix} A_{1,1}(i, i) & A_{1,2}(i, i) & 0 & \dots & 0 \\ A_{2,1}(i, i) & A_{2,2}(i, i) & A_{2,3}(i, i) & \ddots & \vdots \\ 0 & A_{2,3}(i, i) & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & A_{l(i), l(i)-1}(i, i) & A_{l(i), l(i)}(i, i) \end{pmatrix} \quad (4.14)$$

Proof: From the introduction to this theorem it is clear that each stable multivariable linear system has a unique, balanced, representation of this form and that i/o-equivalent systems have exactly the same realization. Therefore this determines a balanced canonical form. It remains to be shown that each system of this form is indeed minimal and stable. This follows from theorem 4.1 together with the fact that the canonical form that is used for the systems with identical singular values has the same property: for each choice of the parameters that is allowed the resulting system is minimal and stable (cf. [4], Section IV.). \square

Remarks:

- i) If $s_1 = p$ then one can just as well parametrize $C_1(i)$ by $C_1(i) = U(i)B(i):U(i)$ a (possibly rectangular) orthogonal matrix.
- ii) Truncation of the last $n - k$ components of the state vector corresponds to the truncation mapping

$$(A, B, C, D) \mapsto ((I_k, 0)A(I_k, 0)^T, (I_k, 0)B, C(I_k, 0)^T, D) \quad (4.17)$$

The canonical form presented has the property that if truncation is applied, the result is again in canonical form. Therefore the resulting lower order system is again minimal and stable.

ACKNOWLEDGMENT

Part of the research for this paper was conducted while the author visited the Dept. Engineering, University of Cambridge and the Center for Engineering Mathematics, University of Texas at Dallas. Discussions with J. Maciejowski, R. Ober, and R. Peeters are gratefully acknowledged.

REFERENCES

- [1] K. Glover, "All optimal Hankel norm approximations of linear multivariable systems and their L^∞ -error bounds," *Int. J. Contr.*, vol. 39, pp. 1115-1193, 1984.
- [2] B. Hanzon, "Identifiability, recursive identification and spaces of linear dynamical systems," CWI Tracts 63, 64, CWI, Amsterdam, 1989.
- [3] B. Hanzon and R. J. Ober, "Overlapping block-balanced canonical forms and parameterizations: The stable SISO case," *SIAM J. Contr. Optim.* submitted.
- [4] B. Hanzon, "Overlapping balanced canonical forms for stable multivariable all-pass systems," in *Challenges of a Generalized System Theory*, P. Dewilde, M. Kaashoek, M. Verhaegen, Eds. North Holland, Amsterdam, 1993, pp. 223-240.
- [5] M. Hazewinkel and R. E. Kalman, "On invariants, canonical forms and moduli for linear, constant finite dimensional dynamical systems," *Lecture Notes in Economics and Mathematical Systems*, vol. 131, 1976, pp. 48-60.
- [6] M. Hazewinkel, "Moduli and canonical forms for linear dynamical systems II: The topological case," *Math. Syst. Th.*, vol. 10, pp. 363-385, 1977.
- [7] P. T. Kabamba, "Balanced forms: canonicity and parametrization," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 1106-1109, 1985.
- [8] J. M. Maciejowski, "Balanced realisations in system identification," in *IFAC Identification and System Parameter Estimation 1985*, H. A. Barker and P. C. Young, Eds. Oxford: Pergamon Press, 1985, pp. 1823-1827.
- [9] R. Ober, "Balanced realizations: Canonical form, parametrization, model reduction," *Int. J. Contr.*, vol. 46, no. 2, pp. 643-670, 1987.
- [10] L. Pernebo and L. M. Silverman, "Model reduction via balanced state space representations," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 382-387, 1982.

Boundary Fractional Derivative Control of the Wave Equation

Brahima Mbodje and Gérard Montseny

Abstract—The wave equation, with fractional derivative feedback at the boundary, is studied. The existence and uniqueness, as well as the asymptotic decay of the solution towards zero is proved. The method used is motivated by the fact that the input-output relationship of generalized diffusion equations, defined on the infinite spatial domain R with collocated sensor and actuator control, can be expressed in terms of fractional integrals. Compared to other methods, the payoff is as follows: 1) The proofs are simpler. 2) The method used can easily be adapted to a wide class of problems involving fractional derivative or integral operators of the time variable.

I. INTRODUCTION

Over the last decade, there has been a considerable amount of research on the modeling and control of beams and waves, spurred on by the increasing use of flexible structures in space technology. In particular, the issue of boundary control has received a special attention. In the case of boundary velocity control, and when the sensor is a deflection sensor, the control law involves first-order time differentiation of signals proportional to the deflection and/or the slope of the deflection at the boundary. Besides it is well known [3], [8] that uniform exponential decay of energy for the wave and the beam equations can be achieved through boundary velocity control. However, when it comes to implementing such controllers, the noise level in the control law could become intolerable due to the differentiation process to be performed. Therefore, to reduce the level of noise, one is naturally led to consider control laws based on lower order derivatives [1], namely fractional order derivative control laws. Another strong motivation for considering fractional derivative controls may be found in [4]. There it is shown, for a Bernoulli-Euler beam, that the H_∞ optimal wave absorbing problem at the boundary is solved by a fractional derivative compensator of order one-half.

Only a few other authors [10], [13] have employed controls involving fractional derivatives to stabilize distributed parameter systems. However, in their investigation, they almost exclusively resort to finite dimensional approximations and techniques pertaining to the Laplace domain.

Our objective is to study via a novel method based on an auxiliary diffusion equation this more general type of feedback control law. As will be noted, the method used here is applicable to other mathematical models such as the Bernoulli-Euler and the Timoshenko beam equations. But in order to keep computational complexities low, we choose to examine the wave equation.

Fractional Calculus

As in [2], we define the fractional derivative operator of order α , $0 < \alpha \leq 1$, by

$$\frac{d^\alpha f}{dt^\alpha}(t) = \int_0^t \frac{(t-\tau)^{-\alpha}}{\Gamma(1-\alpha)} \frac{df}{d\tau}(\tau) d\tau \quad (1)$$

where Γ denotes the well known Gamma function.

Manuscript received July 22, 1993; revised March 28, 1994.

The authors are with the Laboratoire d'Analyse et d'Architecture des Systèmes du CNRS, 31077 Toulouse Cedex, France.

IEEE Log Number 9407002.

As used in (1), the fractional derivative operator is the inverse operator of the fractional integration defined as follows [9]–[12]

$$[I^\alpha f](t) = \int_0^t \frac{(t-\tau)^{\alpha-1}}{\Gamma(\alpha)} f(\tau) d\tau \quad (2)$$

Note that in contrast to the classical derivative operator, the fractional derivative of $f(t)$ does take into account the whole history of the signal $f(t)$ from time 0 up to time t .

From (1) and (2) it is clear that

$$\frac{d}{dt} f = I^{1-\alpha} \left[\frac{d}{dt} f \right] \quad (3)$$

For extension of the above definitions to the case where $\alpha \in \mathbb{R}^+$ the reader is referred to [9], [12].

In the Laplace transform domain, the fractional derivative operator has the property

$$\mathcal{L} \left[\frac{d}{dt} f \right] = s^{-(1-\alpha)} \mathcal{L} \left[\frac{df}{dt} \right] = s^{-\alpha} \mathcal{L}[f] - s^{-(1-\alpha)} f(0) \quad (4)$$

Notice how this equation reduces to the well known formula in the classical case of $\alpha = 1$.

Also, for $0 < \alpha < 1$ and $0 < \beta < 1$ it can be shown by direct calculation that

$$I \left[\frac{d}{dt} f \right] (t) = f(t) - f(0) \quad (5)$$

$$\frac{d}{dt} \left[\frac{d^{\beta} f}{dt^{\beta}} \right] = \frac{d^{\beta+1} f}{dt^{\beta+1}} \quad \text{if } \frac{d^{\beta} f}{dt^{\beta}}(0) = 0 \quad (6)$$

For a more thorough discussion on the properties of fractional calculus see [9]–[12].

Model Definition

We consider the wave plant described by the following equations (without loss of generality the wave velocity can be set to unity)

$$\partial_t \theta(x, t) - \partial_x^2 \theta(x, t) = 0 \quad 0 < x < 1, t \geq 0 \quad (7)$$

$$\begin{aligned} \theta(0, t) &= 0 & \partial_x \theta(1, t) &= -h \partial_t \theta(1, t) \\ 0 &< \alpha < 1, h > 0 \end{aligned} \quad (8)$$

$$\theta(x, 0) = \theta_0(x) \quad \partial_t \theta(x, 0) = \theta_1(x) \quad (9)$$

where h is the constant involved in the boundary control.

Remarks 1) Equations (7)–(9) are clearly a closed loop system with feedback $I^{1-\alpha} \partial_t \theta(1, t)$. 2) Also, the noise amplification caused by the differentiation process at high frequencies is attenuated by the fractional integration.

Our objectives are to prove existence and uniqueness of the solution of system (7)–(9) as well as the asymptotic decay of its energy to zero. Our analysis falls into three parts.

First of all we shall show that the above system can be replaced by an augmented system obtained by coupling the wave equation with a higher order diffusion equation. As will be seen, the main advantage of tackling the problem in this way (rather than in the usual framework of Volterra partial integrodifferential equations) lies in that the analysis of the augmented model will require only a reasonable amount of functional analysis. In fact, the augmented model may be framed into the operator theoretical form $Z(t) = \mathcal{A}Z(t)$ in an appropriate Hilbert space. This very convenient form will then allow the use of results from semigroup theory. More, the appearance of the fractional derivative operator in the second boundary condition in (8) could make the direct analysis of the original set of equations (7)–(9)

much more complicated, as in that case the search for a Lyapunov functional could be more difficult.

The second part of our investigation is devoted to proving existence and uniqueness of the solution of (7)–(9). The third part is concerned with the asymptotic decay of the solution to zero.

II. AUGMENTED SYSTEM

We now demonstrate how a diffusion equation can be used to reformulate system (7)–(9). Consider the higher order diffusion defined on the spatial domain \mathbb{R} with input \mathcal{U}_d

$$\begin{aligned} \partial_t \varphi(y, t) &= (-1)^{l+1} \partial_y^{l+1} \varphi(y, t) \\ &+ \mathcal{U}_d(t) \delta^{(k)}(y) \quad t > 0, y \in \mathbb{A}, k \in \mathbb{N} \end{aligned} \quad (10)$$

$$\varphi(y, 0) = \varphi_0(y) \quad (11)$$

where $\delta^{(k)}$ denotes the k th derivative of the Dirac delta function.

We define the output of the system as follows

$$\mathcal{Y}_l(t) = \langle \delta^{(k)} \varphi(\cdot, t) \rangle \triangleq (-1)^k \partial_y^k \varphi(0, t) \quad (12)$$

Let $f(\xi) = f_{-}^{\infty} f(y) e^{\xi y} dy$ define the Fourier transform of $f(y)$. The following assertion is important in that it constitutes the idea upon which the remainder of the article is based.

Proposition 1 Let $0 \leq k \leq q-1$ and $\varphi(y, 0) = 0$ then setting $\alpha = (2k+1)/2q$ and $\nu = 2q \sin \alpha \pi$ the input–output relationship of system (10)–(12) is given by

$$\mathcal{Y}_l(t) = \nu^{-1} I^{1-\alpha} \mathcal{U}_d(t) \quad (13)$$

Proof It can readily be shown that the self adjoint operator $\mathcal{A} = (-1)^{l+1} \partial_y^{l+1}$ on the Hilbert space $L^2(\mathbb{R})$ generates a contraction semigroup. Also, from the Fourier transform of (10) we have

$$\partial_t \varphi(\xi, t) = -\xi^{2l+1} \varphi(\xi, t) + (-i)^k \xi^k \mathcal{U}_d(t) \quad (14)$$

Hence

$$\varphi(\xi, t) = e^{-\xi^{2l+1} t} \varphi_0(\xi) + (-i)^k \int_0^t \xi^k e^{-\xi^{2l+1} (t-\tau)} \mathcal{U}_d(\tau) d\tau \quad (15)$$

Now setting $\varphi_0(t) = 0$ (zero initial condition) and using (12) we arrive at the following

$$\mathcal{Y}_l(t) = \frac{1}{\pi} \int_0^t \left[\int_0^{\infty} e^{-(t-\tau)\xi^{2l+1}} \xi^k d\xi \right] \mathcal{U}_d(\tau) d\tau \quad (16)$$

$$= \frac{\Gamma(\alpha)}{2q\pi} \int_0^t \frac{1}{(t-\tau)^{1-\alpha}} \mathcal{U}_d(\tau) d\tau \quad \text{where } \alpha = \frac{2k+1}{2q} \quad (17)$$

which is the result looked for since $\Gamma(1-\alpha)\Gamma(\alpha) = \pi[\sin \alpha \pi]^{-1}$. \square

Remarks 1) We realize that by assigning to k and q varying values α will run through all rationals between 0 and 1. Besides, nonrational values of α and the limiting case $\alpha = 1$ can be achieved by merely allowing k to vary continuously within the interval $0 \leq k \leq q-1/2$ which amounts to considering in (10) fractional derivatives of the Dirac distribution [12] as well. 2) Implementation of a control law based on equation (1) is very difficult in the convolution form. On the contrary, system (10)–(12) may be approximated in a very straightforward way by using standard numerical techniques such as finite differences or elements, which lead to very efficient algorithms [7]. These numerical aspects are not developed here but will be the subject of a future paper.

In view of the foregoing proposition, we can model the fractional derivative term in the second boundary condition in (8) by means of

the diffusion system (10)–(12). System (7)–(9) is therefore completely described by

$$\partial_t^2 \theta(x, t) - \partial_x^2 \theta(x, t) = 0 \quad (18)$$

$$\partial_t \varphi(y, t) + (-1)^l \partial_y^{2l} \varphi(y, t) - \partial_t \partial_x \theta(1-t) \delta^{(k)}(y) = 0 \quad (19)$$

$$\theta(0, t) = 0, \quad \partial_x \theta(1, t) = -h \langle \delta^{(k)}, \varphi \rangle \quad (20)$$

$$\begin{aligned} \theta(x, 0) &= \theta_0(x), \quad \partial_t \theta(x, 0) \\ &= \theta_1(x), \quad \varphi(y, 0) = 0 \end{aligned} \quad (21)$$

Remark As will be seen, system (18)–(21) possesses an infinitesimal generator of a C_0 semigroup. In other words, we have managed to replace a system with an hereditary boundary condition by a nonhereditary system.

III. EXISTENCE AND UNIQUENESS

To analyze (18)–(21) we introduce, as in [8], the following notation

$$H_0^1(0, 1) = \{\theta \in L^2(0, 1) : \partial_x \theta \in L^2(0, 1), \theta(0) = 0\}$$

We consider the Hilbert spaces $H_0^1(1, 0) = L^2(0, 1)$ and $L^2(\mathbb{R})$ equipped with the inner products

$$\begin{aligned} (u|v)_{H_0^1(0, 1)} &= \frac{1}{2} \int_0^1 \partial_x u \partial_x v dx = \frac{1}{2} \int_0^1 u v dx \\ (\varphi|\varphi)_{L^2(\mathbb{R})} &= \frac{h}{2i} \int_{\mathbb{R}} \varphi \varphi^* dx \end{aligned} \quad (22)$$

Our state space will be the Hilbert space $\mathcal{H} = H_0^1(0, 1) \times L^2(0, 1) \times L^2(\mathbb{R})$ with inner product

$$\begin{aligned} ((u, v, \varphi)^T | (u, v, \varphi)^T)_{\mathcal{H}} &= \\ &= (u|u)_{H_0^1(0, 1)} + (v|v)_{L^2(0, 1)} + (\varphi|\varphi)_{L^2(\mathbb{R})} \end{aligned} \quad (23)$$

Setting $\Lambda = (\partial_t \partial_x \theta, \varphi)^T \in \mathcal{H}$ (18)–(21) can be put in the abstract form

$$\frac{d}{dt} \Lambda = \Lambda, \quad \Lambda(0) \in \mathcal{H} \quad (24)$$

where $\Lambda: \mathcal{H} \rightarrow \mathcal{H}$ is the linear unbounded operator defined by

$$\Lambda = \begin{bmatrix} 0 & 1 & 0 \\ \partial_x^2 & 0 & 0 \\ 0 & \partial_x \langle \delta_1(x), \cdot \rangle \delta^{(k)} & (-1)^{l+1} \partial_y^{2l} \end{bmatrix} \quad (25)$$

with $\delta_1(x)$ defined by $\langle \delta_1(x), f(x) \rangle = f(1)$. The domain $D(\Lambda)$ of the operator Λ is given by

$$\begin{aligned} D(\Lambda) &= \{(u, v, \varphi)^T \in \mathcal{H} : u \in H_0^2, v \in H_0^1, \\ &\quad v(1) \delta^{(k)} + (-1)^{l+1} \partial_y^{2l} \varphi \in L^2(\mathbb{R}), \\ &\quad \partial_x u(1) + h \langle \delta^{(k)}, \varphi \rangle = 0\} \end{aligned} \quad (26)$$

Note that $(u, v, \varphi)^T \in D(\Lambda)$ implies $\varphi \in H^l(\mathbb{R})$ since, by assumption $0 \leq k \leq q-1$.

We now prove the following proposition about the dissipativity of system (24).

Proposition 2 Let $\mathcal{E}_d(t) = \|\varphi\|_{L^2(\mathbb{R})}^2$ and $\mathcal{E}_u(t) = \|\theta\|_{H_0^1(0, 1)}^2 + \|\partial_t \theta\|_{L^2(0, 1)}^2$ denote the diffusion and the wave energies, respectively. Then the total energy $\mathcal{E}(t) = \mathcal{E}_u(t) + \mathcal{E}_d(t)$ is nonincreasing along the classical solutions of system (24).

Proof Differentiate $\mathcal{E}_i(t)$ with respect to time, use integration by parts and conditions (20)

$$\dot{\mathcal{E}}_i(t) = -h \partial_t \theta(1, t) \langle \delta^{(k)}, \varphi(\cdot, t) \rangle \quad (27)$$

Differentiate $\mathcal{E}_d(t)$ with respect to time, use integrations by parts and the fact that $(\theta, \partial_t \theta, \varphi)^T$, being a classical solution of system (24), φ must belong to $H^l(\mathbb{R})$

$$\dot{\mathcal{E}}_d(t) = -2 \|\partial_y^l \varphi\|_{L^2(\mathbb{R})}^2 + h \partial_t \theta(1, t) \langle \delta^{(k)}, \varphi(\cdot, t) \rangle \quad (28)$$

Now, by adding memberwise (27) and (28), we obtain the desired result

$$\dot{\mathcal{E}}(t) = -2 \|\partial_y^l \varphi\|_{L^2(\mathbb{R})}^2 \quad \square \quad (29)$$

Next we prove the following assertion for another useful estimate.

Proposition 3 Let $\Xi_d(t) = \|\partial_t \varphi\|_{L^2(\mathbb{R})}^2$ and $\Xi_u(t) = \|\partial_t \theta\|_{H_0^1(0, 1)}^2 + \|\partial_t^2 \theta\|_{L^2(0, 1)}^2$, then the sum $\Xi(t) = \Xi_d(t) + \Xi_u(t)$ is increasing along the solutions of (24) with initial conditions in $D(\Lambda)$.

Proof If $\Lambda_0 \in D(\Lambda)$ then $\Lambda(t) = T(t) \Lambda_0$ is a classical solution of (24). Hence we may proceed as in the proof of Proposition 2. We have

$$\dot{\Xi}_d(t) = -h \partial_t \theta(1, t) \langle \delta^{(k)}, \partial_t \varphi(\cdot, t) \rangle \quad (30)$$

and

$$\dot{\Xi}_u(t) = -2 \|\partial_t^l \partial_t \varphi\|_{L^2(\mathbb{R})}^2 + h \partial_t \theta(1, t) \langle \delta^{(k)}, \partial_t \varphi(\cdot, t) \rangle \quad (31)$$

Hence

$$\dot{\Xi}(t) = -2 \|\partial_t^l \partial_t \varphi\|_{L^2(\mathbb{R})}^2 \quad \square \quad (32)$$

We are now able to consider the following

Theorem 1 The operator Λ defined by (25)–(26) generates a C_0 semigroup $\{T(t) : t \geq 0\}$ in \mathcal{H} . For $\Lambda_0 \in D(\Lambda)$ the unique classical solution of (24) is given by $\Lambda(t) = T(t) \Lambda_0$, $\Lambda_0 \in D(\Lambda)$, $t \geq 0$.

Proof By the Lumer-Phillips theorem [11] it suffices to show that Λ is maximal dissipative. Since Proposition 2 already establishes the dissipativity of Λ on the space $D(\Lambda)$, it remains to prove that the operator $\lambda I - \Lambda$ is onto for some $\lambda > 0$. Let $\lambda = 1$. Hence our problem becomes the following

for any $(u, v, \varphi)^T \in \mathcal{H}$, find $(u, v, \varphi)^T \in D(\Lambda)$ such that these three equalities hold

$$\alpha(v) = u(v) - v(v) \quad (33)$$

$$f(v) = v(v) - \partial_x^2 u(v) \quad (34)$$

$$g(y) = \varphi(y) - \partial_t(1) \delta^{(k)}(y) + (-1)^l \partial_y^{2l} \varphi(y) \quad (35)$$

From (33) and (34) and using the fact that $u(0) = v(0) = 0$ it is fairly easily shown that

$$u(v - B) = B \sinh(v) - \int_0^v \{\alpha(\sigma) + f(\sigma)\} \sinh(v - \sigma) d\sigma \quad (36)$$

$$\begin{aligned} v(v - B) &= B \sinh(v) - \alpha(v) - \int_0^v \{\alpha(\sigma) \\ &\quad + f(\sigma)\} \sinh(v - \sigma) d\sigma, \end{aligned} \quad (37)$$

where B is a constant to be determined soon.

Now to show the existence of φ , we first note that

$$\hat{\chi}(\xi) = \frac{\hat{\gamma}(\xi) + (-i\xi)^k \hat{\nu}(1, B)}{1 + \xi^{2q}} \quad (38)$$

solves the Fourier transform of (35) when B is known. Therefore the inverse Fourier transform of $\hat{\chi}$ is the unique solution of (35) and belongs to the space $H^q(\mathbb{R})$. Substituting (36) and the inverse Fourier transform of (38) into the last boundary condition in (26) yields B :

$$\begin{aligned} 2B\pi \cosh(1) - \int_0^1 2\pi \{ \alpha(\sigma) + \beta(\sigma) \} \cosh(1 - \sigma) d\sigma \\ + Bh\nu \sinh(1) \int_{\mathbb{R}} \frac{\xi^{2k}}{1 + \xi^{2q}} d\xi \\ - h\nu \alpha(1) \int_{\mathbb{R}} \frac{\xi^{2k}}{1 + \xi^{2q}} d\xi \\ - h\nu \left\{ \int_0^1 \{ \alpha(\sigma) + \beta(\sigma) \} \sinh(1 - \sigma) d\sigma \right\} \\ \int_{\mathbb{R}} \frac{\xi^{2k}}{1 + \xi^{2q}} d\xi \\ + 2\pi(-1)^k h \int_{\mathbb{R}} \gamma(-y) \partial_y^k \rho(y) dy = 0 \end{aligned} \quad (39)$$

where $\rho(y)$ is the inverse Fourier transform of $[1 + \xi^{2q}]^{-1}$. This ends the proof. \square

IV. ASYMPTOTIC STABILITY

Now consider the subspace of \mathcal{H} defined by

$$\mathcal{S} = \{ z \in \mathcal{H} : \dot{\mathcal{E}}(t) = 0, \forall t \geq 0 \}. \quad (40)$$

Lemma 1: The only classical solution of (24) in the subspace \mathcal{S} is the zero solution.

Proof: Let $(\theta, \partial_t \theta, \varphi)^T \in \mathcal{S}$ be the classical solution of (24); then from (29) we have

$$\forall t \geq 0, \quad \int_{\mathbb{R}} [\partial_y^q \varphi(y, t)]^2 dy = 0 \quad (41)$$

which, together with the fact that $\varphi(\cdot, t) \in L^2(\mathbb{R})$, yields

$$\forall t \geq 0, \quad \varphi(\cdot, t) = 0 \quad \text{a.e. in } \mathbb{R}. \quad (42)$$

From (42) it is clear that system (18)–(20) reduces to the set of equations:

$$\partial_t^2 \theta(x, t) - \partial_x^2 \theta(x, t) = 0, \quad (43)$$

$$\theta(0, t) = 0, \quad \partial_x \theta(1, t) = 0, \quad \partial_t^3 \theta(1, t) = 0. \quad (44)$$

Using (5) and the last equality in (44), we have $\theta(1, t) \equiv \theta(1, 0)$. Combining this with (43)–(44) leads to the identity: $\theta(x, t) \equiv 0$ (In fact, this identity may be obtained by separation of variables). \square

Therefore, according to LaSalle's invariance principle [14], a solution will tend to zero provided its trajectory is precompact in \mathcal{H} . We shall prove the precompactness in Lemma 2 and Proposition 4.

Lemma 2: If $z_0 = (\theta_0, \theta_1, \varphi_0)^T \in D(\Lambda^2)$ then the trajectory of the pair $(\theta, \partial_t \theta)^T$ is compact in $H_0^1(0, 1) \times L^2(0, 1)$.

Proof: First, if $z_0 = (\theta_0, \theta_1, \varphi_0)^T \in D(\Lambda^2)$, then clearly: $(\theta, \partial_t \theta)^T \in H_0^1(0, 1) \times H_0^1(0, 1)$. Now using the differential equation (18) and the fact that $\mathcal{E}(t)$ and $\Xi(t)$ are bounded, we see that

$$\begin{aligned} \|(\theta, \partial_t \theta)^T\|_{H_0^1 \times H_0^1}^2 &\triangleq \int_0^1 |\partial_x \theta|^2 dx \\ &+ \int_0^1 |\partial_x^2 \theta|^2 dx + \int_0^1 |\partial_x \partial_t \theta|^2 dx < C, \quad \text{for some } C > 0. \end{aligned} \quad (45)$$

Hence we conclude by the compact embedding of $H_0^2(0, 1) \times H_0^1(0, 1)$ into $H_0^1(0, 1) \times L^2(0, 1)$. \square

Proposition 4: If $z_0 = (\theta_0, \theta_1, \varphi_0)^T \in D(\Lambda^2)$, then the trajectory of φ , the third component of the solution of (24), is precompact in $L^2(\mathbb{R})$.

Proof: Since for $z_0 \in D(\Lambda^2)$, $\varphi(t)$ is a continuous mapping from $[0, +\infty)$ into $L^2(\mathbb{R})$, it is therefore sufficient to show that $\|\varphi(t)\|_{L^2(\mathbb{R})} \rightarrow 0$ as $t \rightarrow \infty$ whenever $z_0 \in D(\Lambda^2)$.

From relations (29) and (32), we have

$$\int_0^\infty dt \int_{\mathbb{R}} [\partial_t \partial_y^q \varphi]^2 dy < \infty, \quad \int_0^\infty dt \int_{\mathbb{R}} [\partial_y^q \varphi]^2 dy < \infty. \quad (46)$$

Using the above relations together with the Cauchy-Schwartz inequality, we can easily show that

$$\lim_{t \rightarrow \infty} \int_{\mathbb{R}} [\partial_y^q \varphi(y, t)]^2 dy \quad \text{exists.} \quad (47)$$

Now, from the second equality in (46) and from (47), we have

$$\lim_{t \rightarrow \infty} \int_{\mathbb{R}} [\partial_y^q \varphi(y, t)]^2 dy = \lim_{t \rightarrow \infty} \int_{\mathbb{R}} \xi^{2q} |\hat{\varphi}(\xi, t)|^2 d\xi = 0. \quad (48)$$

Therefore, from (48), it is clear that $\|\varphi(t)\|_{L^2(\mathbb{R})}$ will tend to 0, as t approaches ∞ , if

$$\lim_{t \rightarrow \infty} \int_{|\xi| \leq 1} |\hat{\varphi}(\xi, t)|^2 d\xi = 0. \quad (49)$$

We shall now prove the above equality (49) by the dominated convergence theorem.

1) Applying both Parseval's equality and Fubini's theorem to both inequalities in (46), we have

$$\begin{aligned} \int_0^\infty |\partial_t \widehat{\partial_y^q \varphi}(\xi, t)|^2 dt &< \infty, \\ \int_0^\infty |\widehat{\partial_y^q \varphi}(\xi, t)|^2 dt &< \infty, \\ \text{a.e. in } \xi \in [-1, 1]. \end{aligned} \quad (50)$$

So that, by the same arguments that led to (49), we may conclude that

$$\begin{aligned} \lim_{t \rightarrow \infty} |\widehat{\partial_y^q \varphi}(\xi, t)|^2 &= \lim_{t \rightarrow \infty} \xi^{2q} |\hat{\varphi}(\xi, t)|^2 = 0, \\ \text{a.e. in } \xi \in [-1, 1]. \end{aligned} \quad (51)$$

Hence

$$\lim_{t \rightarrow \infty} |\hat{\varphi}(\xi, t)|^2 = 0, \quad \text{a.e. in } \xi \in [-1, 1]. \quad (52)$$

2) Now, taking the Fourier transform in (19), solving for $\hat{\varphi}$ and then using integration by parts with respect to time and the

fact that $\theta(1, t)$ is bounded (Proposition 2), it is not difficult to see that

$$|\varphi(\xi, t)|^2 \leq (|\varphi_0(\xi)|^2 + Q) \in I^1(0, 1) \quad \text{for some } Q > 0 \quad (53)$$

Finally from (52)–(53) and the dominated convergence theorem, we obtain $\lim_{t \rightarrow \infty} \|\varphi(t)\|_{L^2(\mathbb{R})} = 0$ \square

Corollary For $\varphi_0 \in \mathcal{H}$ $\|\varphi(t) - \varphi_0\|_{\mathcal{H}} \rightarrow 0$ as $t \rightarrow \infty$

Proof For $\varphi_0 \in D(\mathcal{A}^2)$ it is a direct consequence of proposition 4, Lemmata 1 and 2 and the LaSalle's invariance principle [14]. Since $D(\mathcal{A}^2)$ is dense in \mathcal{H} the result still holds for $\varphi_0 \in \mathcal{H}$ \square

V. CONCLUSION

The existence, uniqueness and asymptotic decay towards zero of the solution of the coupled wave-diffusion system (24) has been proved. Moreover, since the original system (7)–(9) is embedded in the "augmented" system (24), all the results obtained for the latter carry over to the original system. Also the approach presented here proves to be both simpler and very efficient.

REFERENCES

- [1] R. L. Bagley and R. A. Calico, "Fractional order state equation for the control of viscoelastically damped structures," *J. Guidance*, vol. 14, no. 2, pp. 304–311, Mar–Apr 1991.
- [2] M. Caputo, "Vibrations of an infinite plate with a frequency independent Q ," *J. Acoust. Soc. Amer.*, vol. 60, no. 3, pp. 634–639, Sept 1976.
- [3] G. Chen, "Energy decay estimates and exact boundary value control for the wave equation in a bounded domain," *J. Math. Pures Appl.*, vol. 58, pp. 249–273, 1979.
- [4] M. Khorrami and F. Hironori, "H $_{\infty}$ optimized wave absorbing control: Analytical and experimental results," *J. Guidance Contr. Dynamics*, vol. 16, no. 6, p. 1146, Nov–Dec 1993.
- [5] J. F. Lagnese, "Boundary stabilization of thin plates," *SIAM Studies Appl. Math.*, vol. 10, 1989.
- [6] G. Montseny, J. Audouinet, and B. Mbodje, "A simple viscoelastic damper model: Application to vibrating string," presented at the 10th Int. Conf. Analysis and Optimization of Systems, INRIA, Rocquencourt, France, June 1992.
- [7] —, "Optimal models of fractional integrators—Application to viscoelastic systems," presented at the IEEE/SMC Conf. 1c, Touquet, France, Oct 1993, invited paper.
- [8] Ö. Morgül, "Dynamic boundary control of a Euler–Bernoulli beam," *IEEE Trans. Automat. Contr.*, vol. 37, no. 5, pp. 639–642, May 1992.
- [9] K. B. Oldham and J. Spanier, *The Fractional Calculus: Theory and Application of Differentiation and Integration to Arbitrary Order*. Orlando, FL: Academic Press, 1974.
- [10] A. Oustaloup, *La commande CRONE*. Paris: Hermès, 1991.
- [11] A. Pazy, *Semigroup of Linear Operators and Applications to Partial Differential Equations*. New York: Springer-Verlag, 1983.
- [12] G. F. Shilov, *Generalized Functions and Partial Differential Equations*. New York: Gordon and Breach, 1968.
- [13] S. B. Skaer, A. N. Michel, and R. K. Miller, "Stability of viscoelastic control systems," *IEEE Trans. Automat. Contr.*, vol. 33, no. 4, pp. 348–357, Apr 1988.
- [14] A. J. Walter, *Dynamical Systems and Evolution Equations: Theory and Applications*. New York: Plenum Press, 1980.

Observer-Based Parameter Identifiers for Nonlinear Systems with Parameter Dependencies

Shahab Sheikholeslam

Abstract—The class of nonlinear dynamical systems whose dynamics depends linearly on the unknown parameter vector is considered. After reviewing a standard observer-based identifier for estimating the unknown parameters, we propose a family of new identifiers which exploit the *a priori* known parameter dependencies. Then we establish that, under mild assumptions on the dynamical system, 1) the proposed identifiers are stable, 2) the weighted norm of state-parameter errors using the proposed identifiers are less than the corresponding errors using the standard identifier for a length of time after $t = 0$. The main contribution of this paper is that it introduces a family of observer-based identifiers which exhibit better transient performance than the standard identifier.

I. INTRODUCTION

The subject of adaptive identification and control has been studied extensively in the literature. For the basic theory refer to [8]–[5], [11]–[9]. Adaptive parameter identifiers studied in the literature fall into two classes: 1) identification algorithms which use the input–output relations to continuously estimate the unknown parameters (see [8]–[5]) and 2) model reference identification algorithms [11]–[9] which use an observer to continuously estimate the unknown parameters. The algorithms in class 2) are useful for adaptive control. The identifiers in this paper belong to class 2).

In most engineering applications the unknown parameters in the system equations are known functions of the physical parameters of the model. In many such cases the unknown parameters are known functions of a strictly proper subset of all the unknown parameters. Our approach in designing new parameter identifiers for a class of nonlinear systems is to exploit such parameter dependencies while maintaining the stability of the identifier. Intuitively an identifier which exploits the *a priori* known parameter dependencies should yield smaller errors than a similar identifier which does not make use of such dependencies. In [3] and [1] the authors propose identification algorithms in class 1) that make use of parameter dependencies. The main contribution of this paper is that it introduces a family of identifiers in class 2) that exhibit better transient performance than a standard identifier. The method of proof used to show the superior transient performance can be used to prove that the parameter errors resulting from using the identifier in [3] in class 1) are smaller than the corresponding errors using a standard least squares identification algorithm.

Recently some authors have studied the problem of robust adaptive control of linear time invariant systems [6]–[4]. As pointed out in [6] one of the main limitations to the theory is the lack of a more convenient quantification of the transient performance. We believe that the results presented in this paper are preliminary attempts at such a quantification.

The paper is organized as follows. Section II proposes a new observer-based identifier for a class of nonlinear dynamical systems and states theorems regarding its stability and transient performance to compare the transient performance of the proposed identifier with

Manuscript received July 1992; revised January 1993. This work was supported in part by the PATH Project under Grant RTA 74H221.

The author is with General Electric Company R&D, Schenectady, NY 12301 USA.

IEEE Log Number 9407001.

that of the standard identifier, Section III shows simulation results for adaptive identification of the longitudinal dynamics of a vehicle; Section IV outlines the main results and provides future directions for research; Section V contains the proofs of the theorems.

II. PARAMETER IDENTIFIERS

Throughout this paper, we use $|\cdot|$ to denote the Euclidean norm of a vector and $\|\cdot\|$ to denote the induced norm of a matrix.

A. Background

Consider the nonlinear time-invariant dynamical system

$$\begin{aligned}\dot{x} &= W^T(x, u)\theta^* \\ x(0) &= x_0\end{aligned}\quad (2.1)$$

where $x \in R^n$ denote the state of the system, $u \in R^m$ denotes the control input to the system, $\theta^* \in R^p$ is a constant vector of unknown parameters, and $W(\cdot, \cdot)$ is a $p \times n$ matrix of piecewise continuous functions. W^T denotes the transpose of the matrix W .

Parameter Structure: In (2.1), components of the unknown parameter vector θ^* are known functions of the physical parameters (e.g., engine time-lag (τ) and drag coefficient (K_d) in the vehicle dynamics (2.10) with θ^* defined in (2.13)). Previous papers [13], [14], [16], [2] have proposed observer-based identifiers, for (2.1), which do not exploit these parameter dependencies.

Identifier Structure (Not Using Parameter Dependencies) [13], [14], [16], [10]: Let $A \in R^{n \times n}$ be a Hurwitz matrix and $Q \in R^{n \times n}$ be a given symmetric positive definite matrix; let $P \in R^{n \times n}$ denote the unique symmetric positive-definite solution of the Lyapunov equation

$$A^T P + P A = -Q. \quad (2.2)$$

For the dynamical system in (2.1) the identifier is chosen to be

$$\dot{\hat{x}}'' = A(\hat{x}'' - x) + W^T(x, u)\hat{\theta}'' \quad (2.3)$$

$$\dot{\hat{\theta}}'' = -W(x, u)P(\hat{x}'' - x) \quad (2.4)$$

$$\hat{x}''(0) = \hat{x}_0 \quad (2.5)$$

$$\hat{\theta}''(0) = \hat{\theta}_0. \quad (2.6)$$

The superscript u in \hat{x}'' , $\hat{\theta}''$ indicates that the Lyapunov function, used to prove the stability of the above identifier, is unconstrained with respect to parameter dependencies in (2.1).

Assumption 1 (Boundedness of Regressor): We assume that $W: R^n \times R^m \rightarrow R^{p \times n}$ is bounded.

Stability of the Identifier: Define the state error and the parameter error by

$$e'' = \hat{x}'' - x \quad (2.7)$$

$$\phi'' = \hat{\theta}'' - \theta^*. \quad (2.8)$$

Theorem 1 (Stability of Identifier (2.3)–(2.4)): Consider the dynamical system (2.1) with the identifier (2.3)–(2.4). Let Assumption 1 hold; then $e'' \in L_2 \cap L_\infty$, $\dot{e}'' \in L_\infty$, and $\phi'' \in L_\infty$.

Proof of Theorem 1: See [16 (and references therein)], [11, sec. 2.4]. The proof of Theorem 1 uses a standard unconstrained Lyapunov function candidate

$$V''(e'', \phi'') = \frac{1}{2} e''^T P e'' + \frac{1}{2} \phi''^T \phi''. \quad (2.9)$$

B. New Identifier

In many engineering applications, the components of $\theta^* \in R^p$ are known functions of a strictly proper subset of θ_i^* for $i = 1, 2, \dots, p$. (θ_i^* denotes the i th component of θ^* .) This is illustrated in the following example.

Example: Consider the following nonlinear differential equation representing the longitudinal dynamics of a vehicle [15], [12]:

$$\ddot{x} = -2 \frac{K_d}{m} \dot{x} \ddot{x} - \frac{1}{\tau} \left[\dot{x} + \frac{K_d}{m} \dot{x}^2 + \frac{d_m}{m} \right] + \frac{u}{m\tau} \quad (2.10)$$

where x is the position of the vehicle with respect to a fixed reference point O on the road; K_d denotes the vehicle's aerodynamic drag coefficient; m denotes the mass of the vehicle; τ denotes the vehicle's engine time-constant; d_m denotes the mechanical drag of the vehicle; and u denotes the throttle-command input to the vehicle's engine.

Equation (2.10) can be put into the form

$$\ddot{x} = w^T(\dot{x}, \ddot{x}, u)\theta^* \quad (2.11)$$

where

$$w(\dot{x}, \ddot{x}, u) = \left[-\frac{2}{m} \dot{x} \ddot{x}, -\ddot{x} - \frac{d_m}{m} + \frac{u}{m}, -\frac{\dot{x}^2}{m} \right]^T \quad (2.12)$$

and

$$\theta^* = \left[K_d, \frac{1}{\tau}, \frac{K_d}{\tau} \right]^T \quad (2.13)$$

For the vehicle dynamics (2.11), $\theta_1^* = \theta_2^* \theta_3^*$.

Our approach in designing a new parameter identifier for (2.1) is to exploit such parameter dependencies while maintaining the stability of the identifier.

Parameter Dependencies:

Definition (Parameter Constraint Set): Consider the dynamical system (2.1) with $\theta^* \in R^p$. The parameter constraint set is an N -dimensional C^2 submanifold of R^p parameterized by $\theta_1^*, \dots, \theta_N^*$. Hence, the parameter constraint manifold is described by

$$\begin{aligned}C(\{\gamma_i, i = N+1, \dots, p\}) \\ = \{\theta \in R^p \mid \theta_i = \gamma_i(\theta_1, \dots, \theta_N) \text{ for } i = N+1, \dots, p\}\end{aligned} \quad (2.14)$$

where for $i = N+1, \dots, p$, $\gamma_i: R^N \rightarrow R$ is a known C^2 map (i.e., γ_i is at least twice continuously differentiable).

Penalty function: Consider the parameter constraint set in (2.14). We define a penalty function $\hat{P}: R^p \rightarrow R$ with

$$\hat{P}(\hat{\theta}) = \frac{1}{2} \sum_{i=N+1}^p \lambda_i [\hat{\theta}_i - \gamma_i(\hat{\theta}_1, \dots, \hat{\theta}_N)]^2 \quad (2.15)$$

where $\lambda_i > 0$ for all $i = N+1, \dots, p$.

For the vehicle dynamics (2.11), the parameter constraint set is $C(\{\gamma_3\}) = \{\theta \in R^3 \mid \theta_3 = \gamma_3(\theta_1, \theta_2) = \theta_1 \theta_2\}$. The penalty function (2.15) with respect to the parameter constraint set $C(\{\gamma_3\})$ is

$$\hat{P}(\hat{\theta}) = \frac{1}{2} \lambda_3 [\hat{\theta}_3 - \hat{\theta}_1 \hat{\theta}_2]^2 \quad (2.16)$$

where $\lambda_3 > 0$.

Identifier Structure (Using Parameter Dependencies): Let $A \in R^{n \times n}$, $Q \in R^{n \times n}$, and $P \in R^{n \times n}$ be the same matrices used in (2.2) and (2.3)–(2.4). Let \hat{P} be the penalty function in (2.15) for the parameter constraint set in (2.14).

The new identifier for the dynamical system (2.1) is

$$\dot{\hat{x}}^c = A(\hat{x}^c - x) + W^T(x, u)[\hat{\theta}^c + D\hat{P}(\hat{\theta}^c)] \quad (2.17)$$

$$\dot{\hat{\theta}}^c = -W(x, u)P(\hat{x}^c - x) \quad (2.18)$$

$$\hat{x}^c(0) = \hat{x}_0 \quad (2.19)$$

$$\hat{\theta}^c(0) = \hat{\theta}_0 \quad (2.20)$$

where $D\hat{P}(\cdot)$ denotes the differential of the penalty function.

We have used the superscript c in \hat{x}^c , $\hat{\theta}^c$ to indicate that the Lyapunov function candidate, used to prove the stability of the above identifier, is constrained with respect to parameter dependencies in (2.1). (See (2.23) below.)

Assumption 2 (Existence and Uniqueness): Throughout this paper, we assume that the solutions of the differential equations (2.17)–(2.18) with initial conditions (2.19)–(2.20) exist and are unique on $[0, \infty)$.

Stability of the Identifier: Denote

$$\epsilon' := \hat{x}' - x \quad (2.21)$$

$$\phi' := \hat{\theta}' - \theta^*. \quad (2.22)$$

Theorem 2 (Stability of Identifier (2.17)–(2.18)): Consider the dynamical system (2.1) with the identifier (2.17)–(2.18). Let ϵ' , ϕ' denote the identifier's state-error and parameter error vectors, respectively. Suppose Assumption 1 (Boundedness of Regressor) and Assumption 2 (Existence and Uniqueness) hold, then $\epsilon' \in L_2 \cap L_\infty$, $\dot{\epsilon}' \in L_\infty$, and $\phi' \in L_\infty$.

Proof (Theorem 2) See the Appendix.

Comments: a) Proof of Theorem 2 uses a Lyapunov function candidate of the form

$$V'(\epsilon', \phi') := \frac{1}{2} \epsilon'^T P \epsilon' + \frac{1}{2} \phi'^T \phi' + P(\phi' + \theta^*). \quad (2.23)$$

Comparing (2.9) with (2.23), we note that the penalty term $\hat{P}(\phi' + \theta^*)$ is the new addition to the standard Lyapunov function candidate. The design of the new identifier (2.17)–(2.18) is motivated by using the Lyapunov function candidate (2.23).

b) In [3, (2.14)], the authors propose a similar Lyapunov function for improving the robustness of a least squares type identification algorithm. The main differences between the identification algorithms presented in [3] and the observer-based identifier in (2.17)–(2.18) are: 1) the identifier in (2.17)–(2.18) belongs to the so-called model reference identification family of algorithms [11, p. 50]. The algorithms in this family are useful for adaptive control. The identification algorithms in [3] use the input-output relations to estimate the unknown parameters; 2) the dynamical systems in [3] are SISO, linear, time-invariant, whereas, we consider MIMO, nonlinear, time-invariant dynamical systems (see (2.1)); and 3) in [3], the parameters are multilinear functions of the unknowns (i.e., the γ_i , see (2.14), are multilinear functions); whereas, we allow more general parameter dependencies in the definition of the parameter constraint set (2.14).

Transient Performance We now compare the transient performance of the new identifier (2.17)–(2.18) with that of the standard identifier (2.3)–(2.4).

Consider the dynamic system represented by (2.1) with the parameter constraint set (2.14). Let $t \mapsto \theta''(t) := (\theta_1''(t), \dots, \theta_N''(t), \hat{\theta}_{N+1}''(t), \dots, \hat{\theta}_p''(t))^T \in R^n$ denote the solution of (2.4) with initial condition $\theta''(0)$.

Assumption 3 (Deviation from Constraint Set) We assume that there exists a $j \in \{N+1, \dots, p\}$ such that

$$\left. \frac{d}{dt} [\hat{\theta}_j''(0) - \gamma_j(\hat{\theta}_1''(0), \dots, \hat{\theta}_N''(0))] \right|_{t=0} \neq 0. \quad (2.24)$$

Remark. Intuitively, Assumption 3 indicates that the direction of the velocity vector of the parameter estimates at $t = 0$, using the update law (2.4), lies outside the tangent space of the parameter constraint set (2.14) at $\theta''(0)$. Hence, after $t = 0$, the parameter estimates may not lie on the parameter constraint set (2.14).

Theorem 3 (Transient Performance): Consider the dynamical system represented by (2.1). Let the proposed identifier (2.17)–(2.18) and the standard identifier (2.3)–(2.4) have the same initial conditions (2.5)–(2.6). Suppose the parameter estimates initially lie on the parameter constraint set (2.14). Suppose Assumption 1 (Boundedness of Regressor), Assumption 2 (Existence and Uniqueness), and Assumption 3 (Deviation from Constraint Set) hold. Let $\epsilon'' : R_+ \rightarrow R^n$, $\phi'' : R_+ \rightarrow R^n$ be defined as in (2.7), (2.8), respectively. Let

$\epsilon' : R_+ \rightarrow R^n$, $\phi' : R_+ \rightarrow R^n$ be defined as in (2.21), (2.22), respectively. Under these conditions, there exists an $\hat{\epsilon} > 0$ such that for all $\epsilon \in [0, \hat{\epsilon})$,

$$\epsilon'^T(\epsilon) P \epsilon'(\epsilon) + \phi'^T(\epsilon) \phi'(\epsilon) \leq \epsilon''^T(\epsilon) P \epsilon''(\epsilon) + \phi''^T(\epsilon) \phi''(\epsilon). \quad (2.25)$$

Proof (Theorem 3) See the Appendix.

Comments 1) Theorem 3 is the main result of this paper. It shows that initially the proposed identifier (2.17)–(2.18) results in smaller state and parameter errors than the standard identifier (2.3)–(2.4). In [3], the authors present simulation results to verify the improvement in the overall robustness characteristics of their proposed identifiers; they do not prove that their identifiers result in smaller parameter errors.

2) Theorem 3 does not make any assertions comparing the magnitude of the parameter errors between the two identifiers. Using the same steps in the proof of Theorem 3 for the least squares type identification algorithm in [3, eq. (2.18)], one can show that

there exists an $\hat{\epsilon} > 0$ such that for all $\epsilon \in [0, \hat{\epsilon})$,

$$\phi'^T(\epsilon) \phi'(\epsilon) \leq \phi''^T(\epsilon) \phi''(\epsilon). \quad (2.26)$$

Whether (2.26) remains valid for the identifier (2.17)–(2.18) remains an open question.

3) In general the largest value of ϵ , for which (2.25) holds, depends on the choice of the penalty function $\hat{P}(\cdot)$, the exogenous input vector u , and the dynamics of the system.

4) At the present time, there are no systematic procedures for selecting the values of λ_i in (2.15). Intuitively, the choice of the λ_i reflects a tradeoff between the weighted norm of the identifier's state-parameter error vectors and the deviation of the parameter estimates from the parameter constraint set.

III SIMULATION

To compare the transient performance of the proposed identifier (2.17)–(2.18) with that of the standard identifier (2.3)–(2.4), we ran simulations using the nonlinear differential equation (2.10) representing the longitudinal dynamics of a vehicle.

The standard identifier for (2.11) is

$$\frac{d}{dt} \dot{x}'' = -\sigma_x (\ddot{x}'' - \ddot{x}) + w^T(\dot{x}, \ddot{x}, u) \theta'' \quad (3.1)$$

$$\frac{d}{dt} \theta'' = -\sigma_\theta w(\dot{x}, \ddot{x}, u) (\dot{x}'' - \dot{x}) \quad (3.2)$$

where $\dot{x}''(0) = \dot{x}_0$, $\theta''(0) = \theta_0$, $\ddot{x}'' \in R$, $\theta'' \in R^1$, $\sigma_x > 0$, and $\sigma_\theta > 0$.

Using the penalty term $P(\cdot)$ in (2.16), the proposed identifier for (2.11) is

$$\frac{d}{dt} \dot{x}' = -\sigma_x (\dot{x}' - \dot{x}) + w^T(\dot{x}, \ddot{x}, u) [\hat{\theta}' + D \hat{P}(\theta')] \quad (3.3)$$

$$\frac{d}{dt} \theta' = -\sigma_\theta w(\dot{x}, \ddot{x}, u) (\dot{x}' - \dot{x}) \quad (3.4)$$

where $\dot{x}'(0) = \dot{x}_0$, $\theta'(0) = \hat{\theta}_0$, $\dot{x}' \in R$, and $\theta' \in R^1$.

Vehicle Parameters In all the simulations conducted, the vehicle parameters were as follows: $m = 1136$ kg, $d_m = 150.74$ N, $K_d = 0.396$ kg/m, and $\tau = 0.18$ s.

Identifier Parameters We chose $\sigma_x = 1$ and $\sigma_\theta = 2$. For the proposed identifier (3.3)–(3.4), $\lambda_1 = 2$ was used for the penalty term $P(\theta)$ in (2.16). The initial state and parameter estimates were $\dot{x}_0 = 1$, $\hat{\theta}_0 = [0.44, 5, 2.2]^T$, respectively.

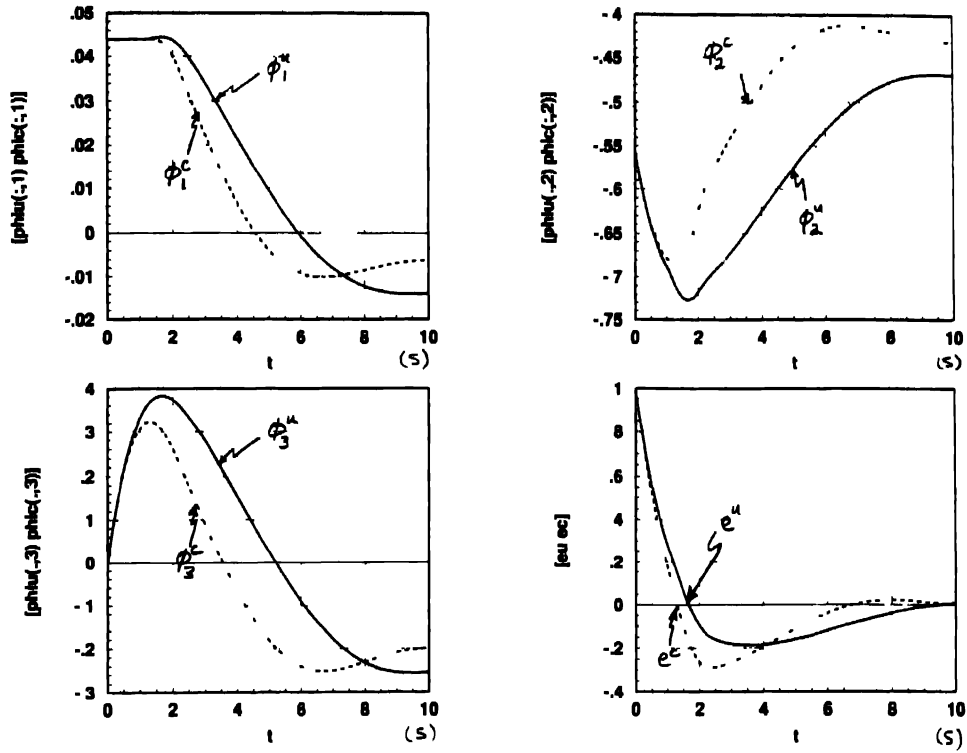


Fig. 1 $\phi_i(t)$, $\hat{\phi}_i(t)$ ($i = 1, 2, 3$), $e(t)$ and $\hat{e}(t)$ versus t

Simulation Setup Initially the vehicle was traveling at a constant speed of 17.9 m/s (i.e. $v(0) = 17.9$ m/s, $\dot{v}(0) = 0$ m/s², $\ddot{v}(0) = 0$ m/s³). Starting at time $t = 1$ s, the throttle command input increased linearly from its initial steady state value of 381 N at a rate of 1136 N/s until it reached its final value of 1517 N at $t = 2$ s.

Simulation Results Fig. 1 shows the state and the parameter errors for the standard identifier (3.1)–(3.2) and the proposed identifier (3.3)–(3.4). Note that the magnitude of the parameter errors using the proposed identifier (3.3)–(3.4) are smaller than the corresponding errors using the standard identifier (3.1)–(3.2) for a length of time after $t = 0$ s. Also note that in either case the parameter errors do not converge to zero.

Fig. 2 shows the weighted norm of the state parameter error vectors for the standard identifier (3.1)–(3.2) (i.e. $t \mapsto \sqrt{\epsilon'(t)\epsilon(t) + \phi'(t)\phi(t)}$) and the proposed identifier (3.3)–(3.4) (i.e. $t \mapsto \sqrt{\epsilon(t)\epsilon(t) + \phi(t)\phi(t)}$). Note that for all $t \in [0, 10]$, $\sqrt{\epsilon(t)\epsilon(t) + \phi(t)\phi(t)} \leq \sqrt{\epsilon'(t)\epsilon(t) + \phi'(t)\phi(t)}$.

These simulation results show that the proposed identifier (3.3)–(3.4) has better transient performance than the standard one (3.1)–(3.2). Thus, we expect that using the parameter estimates from the identifier (3.3)–(3.4) together with a certainty-equivalence control law will result in superior closed-loop adaptive-control performance of a vehicle's longitudinal dynamics. Simulations of indirect adaptive control laws for a platoon of vehicles support this conclusion.

IV. CONCLUSION

We have proposed a family of observer-based identifiers for a class of nonlinear dynamical systems which use the a priori known parameter dependencies. Under mild assumptions on the dynamical system, we have shown that 1) the proposed identifiers are stable (see Theorem 2), and 2) the weighted norm of state-parameter errors using the new identifier are less than the corresponding errors using the standard identifier, for a length of time after $t = 0$ (see Theorem 3).

There remain a number of open questions:

1) How can we select the coefficients λ_i for $i = \lambda + 1, \dots, p$ in the definition of the penalty function $P(\cdot)$ (see (2.15)) so as to attain the 'best' transient performance?

2) Is the proposed identifier more robust with respect to parameter variations and exogenous measurement noise?

3) Can we design stable identifiers which have the least number of states (e.g. the update law only updates the basic parameters; other parameter estimates are then computed using estimates of the basic parameters)?

We believe that satisfactory answers to the above questions will lead to systematic strategies for designing robust nonlinear adaptive control laws.

APPENDIX

Proof of Theorem 2

Consider the following Lyapunov function candidate

$$V(\epsilon, \phi) = \frac{1}{2}\epsilon^T P \epsilon + \frac{1}{2}\phi^T \phi + P(\phi + \theta^*) \quad (5.1)$$

Differentiating both sides of (5.1) with respect to time, and using (2.2), we get

$$\dot{V}(\epsilon(t), \phi(t)) = -\frac{1}{2}\epsilon^T(t)Q\epsilon(t) \leq 0 \quad (5.2)$$

Hence, by the standard reasoning, $\epsilon \in L_\infty$ and $\phi \in L_\infty$. Since V is bounded (i.e., Assumption 1), $DP(\cdot)$ is bounded ($DP(\cdot)$ is continuous and $\theta = \phi + \theta^* \in L_\infty$), $\epsilon \in L_\infty$ and $\phi \in L_\infty$, we get $\epsilon \in L_\infty$.

Integrating both sides of (5.2) from zero to t and noting that $Q = Q^T \geq \lambda_1^Q I_{n \times n} > 0$ (λ_1^Q is the smallest eigenvalue of Q), we get

$$V(\epsilon'(t), \phi(t)) - V(\epsilon(0), \phi(0)) \leq -\frac{1}{2}\lambda_1^Q \int_0^t \epsilon^T(\tau)\epsilon(\tau) d\tau \quad (5.3)$$

Dividing both sides of (5.3) by $-\frac{1}{2}\lambda_1^Q$, we get $\epsilon \in L_2$. \square

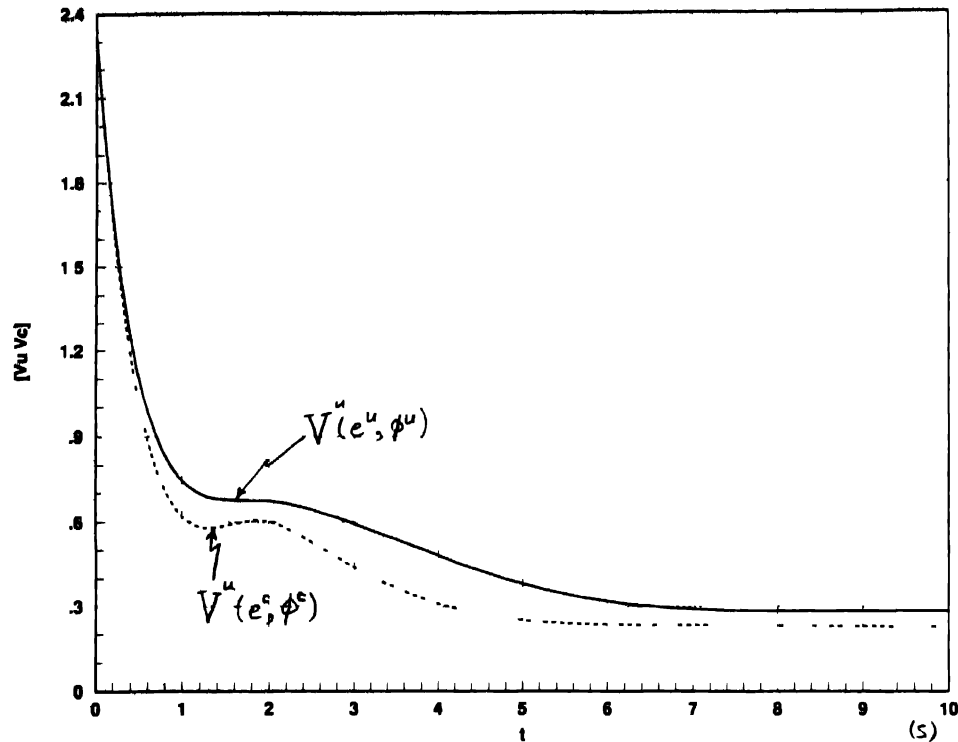


Fig. 2 $V(u(t), v(t)) - V(u(t), v'(t))$ versus t

Proof of Theorem 3

Consider the Lyapunov function candidates (2.9) and (5.1) for the error dynamics (2.3), (2.4) and (2.17), (2.18) respectively. Since the initial parameter estimate θ_0 lies on the parameter constraint set (2.14) (i.e., $P(\theta_0) = 0$) and both identifiers have the same initial conditions from (2.9) and (5.1) we get

$$V(u(t), v(t)) = V(u(t), v'(t)) \quad (5.4)$$

Differentiating both sides of (2.9) with respect to time we get

$$V_{(2.17)(2.18)}(u(t), v(t)) = -\frac{1}{2} \epsilon^{-1}(t) Q \epsilon(t) \quad (5.5)$$

Thus from (5.2), (5.5) and given $\epsilon(0) = \epsilon'(0)$ we get

$$V_{(2.17)(2.18)}(u(0), v(0)) = V_{(2.17)(2.18)}(u(0), v'(0)) \quad (5.6)$$

Since $\epsilon(0) = \epsilon'(0)$, $\epsilon(0) = \epsilon'(0)$, $DP(\epsilon(0) + \theta^*) = 0$

$$\epsilon(0) = \epsilon'(0) \quad (5.7)$$

Differentiating both sides of (5.5) and (5.2) with respect to time (5.7) and given $\epsilon(0) = \epsilon'(0)$ we get

$$V_{(2.17)(2.18)}(u(0), v(0)) = V_{(2.17)(2.18)}(u(0), v'(0)) \quad (5.8)$$

Writing the Taylor expansion of $V(u(\epsilon), v(\epsilon)) - V(u(\epsilon'), v(\epsilon'))$ about $t = 0$ and noting (5.4), (5.6) and (5.8) we get

$$V(u(\epsilon), v(\epsilon)) - V(u(\epsilon'), v(\epsilon')) = O(\epsilon^4) \quad (5.9)$$

where $\epsilon \mapsto O(\epsilon^4)$ denotes a function such that $\lim_{\epsilon \rightarrow 0} |O(\epsilon^4)/\epsilon^4| < \infty$

For $i = N+1, \dots, p$ let

$$q_i(t) = \theta_i(t) - \mu_i(\theta_1(t), \dots, \theta_N(t)) \quad (5.10)$$

Since $\theta(0)$ lies on the parameter constraint set (2.14), for $i = N+1, \dots, p$ we have

$$q_i(0) = 0 \quad (5.11)$$

Since $\epsilon(0) = \epsilon'(0)$, $\epsilon(0) = \epsilon'(0)$ (for $i = N+1, \dots, p$)

$$q_i(0) = \frac{d}{dt} [\theta_i(0) - \mu_i(\theta_1(0), \dots, \theta_N(0))] \quad (5.12)$$

Successively differentiating both sides of (2.15) with respect to time along the solution trajectories of (2.17), (2.18) we get

$$P_{(2.17)(2.18)}(\theta) = \sum_{j=N+1}^i \lambda_j q_j \quad (5.13)$$

$$P_{(2.17)(2.18)}(\theta) = \sum_{j=N+1}^i \lambda_j (q_j + q_j q_j) \quad (5.14)$$

Hence from (5.11), (2.15), (5.13) and (5.14) we get

$$P(\theta(0)) = 0 \quad (5.15)$$

$$P_{(2.17)(2.18)}(\theta(0)) = 0 \quad (5.16)$$

$$P_{(2.17)(2.18)}(\theta(0)) = \sum_{j=N+1}^i \lambda_j q_j(0) \quad (5.17)$$

From (5.12) and Assumption 3 (deviation from constraint set) we note that for some $j \in \{N+1, \dots, p\}$, $q_j(0) \neq 0$. Hence from (5.17) we note that

$$P_{(2.17)(2.18)}(\theta(0)) \geq \lambda_j q_j^*(0) > 0 \quad (5.18)$$

Writing the Taylor expansion of $P(\theta(\epsilon))$ about $t = 0$ and noting (5.15) and (5.16), we get

$$P(\theta(\epsilon)) = \frac{\epsilon^2}{2} P_{(2.17)(2.18)}(\theta(0)) + O(\epsilon^4) \quad (5.19)$$

From the definition of V' in (5.1) and V'' in (2.9) we note that

$$V(u(\epsilon), v(\epsilon)) - V(u(\epsilon'), v(\epsilon')) = V''(u(\epsilon), v(\epsilon)) + P(\epsilon(\epsilon) + \theta^*) \quad (5.20)$$

Substituting the expression for $V^c(e^c(\epsilon), \phi^c(\epsilon))$ from (5.20) into (5.9), we get

$$V^u(e^c(\epsilon), \phi^c(\epsilon)) + \tilde{P}(\hat{\theta}^c(\epsilon)) - V^u(e^u(\epsilon), \phi^u(\epsilon)) = O(\epsilon^3). \quad (5.21)$$

Substituting the expression for $\tilde{P}(\hat{\theta}^c(\epsilon))$ from (5.19) into (5.21), after algebraic manipulations, we get

$$\begin{aligned} V^u(e^c(\epsilon), \phi^c(\epsilon)) - V^u(e^u(\epsilon), \phi^u(\epsilon)) \\ = -\frac{\epsilon^2}{2} \ddot{P}_{(2,17)(2,18)}(\hat{\theta}^c(0)) + O(\epsilon^3). \end{aligned} \quad (5.22)$$

From (5.18) and (5.22) we get

$$\begin{aligned} V^u(e^c(\epsilon), \phi^c(\epsilon)) - V^u(e^u(\epsilon), \phi^u(\epsilon)) \\ \leq -\frac{\epsilon^2}{2} \lambda_j \dot{q}_j^2(0) + O(\epsilon^3). \end{aligned} \quad (5.23)$$

By definition, $\lim_{\epsilon \rightarrow 0} |O(\epsilon^3)/\epsilon^3| < \infty$; hence, there exists an $\bar{\epsilon} > 0$ such that for all $\epsilon \in [0, \bar{\epsilon}]$, the right hand side of (5.23) is negative, i.e.,

$$-\frac{\epsilon^2}{2} \lambda_j \dot{q}_j^2(0) + O(\epsilon^3) \leq 0. \quad (5.24)$$

Thus, from (5.23) and (5.24) we get, for all $\epsilon \in [0, \bar{\epsilon}]$,

$$V^u(e^c(\epsilon), \phi^c(\epsilon)) - V^u(e^u(\epsilon), \phi^u(\epsilon)) \leq 0. \quad (5.25)$$

Noting the definition of V^u in (2.9), from (5.25) we get, for all $\epsilon \in [0, \bar{\epsilon}]$,

$$e^{cT}(\epsilon) P e^c(\epsilon) + \phi^{cT}(\epsilon) \phi^c(\epsilon) \leq e^{uT}(\epsilon) P e^u(\epsilon) + \phi^{uT}(\epsilon) \phi^u(\epsilon). \quad (5.26)$$

ACKNOWLEDGMENT

The author would like to thank Prof. C. A. Desoer for his useful suggestions.

REFERENCES

- [1] G. Bastin, R. R. Bitmead, G. Campion, and M. Gevers, "Identification of linearly over-parameterized nonlinear systems," in *Proc. Conf. Decision and Control*, 1989, pp. 618-623.
- [2] G. Bastin and M. R. Gevers, "Stable adaptive observers for nonlinear time-varying systems," *IEEE Trans. Automat. Contr.*, vol. 33, no. 7, pp. 650-658, July 1988.
- [3] S. Dasgupta, B. D. O. Anderson, and R. J. Kaye, "Identification of physical parameters in structured systems," *Automatica*, vol. 24, no. 2, pp. 217-225, Mar. 1988.
- [4] F. Giri, M. M'Saad, L. Dugard, and J. M. Dion, "Robust adaptive regulation with minimal prior knowledge," *IEEE Trans. Automat. Contr.*, vol. 37, no. 3, pp. 305-315, Mar. 1992.
- [5] G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
- [6] J. M. Krause, P. P. Khargonekar, and G. Stein, "Robust adaptive control: Stability and asymptotic performance," *IEEE Trans. Automat. Contr.*, vol. 37, no. 3, pp. 316-331, Mar. 1992.
- [7] G. Kreisselmeier, "Adaptive observers with exponential rate of convergence," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 2-8, 1977.
- [8] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
- [9] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [10] J. Pomet and L. Praly, "Indirect adaptive nonlinear control," in *Proc. Conf. Decision and Control*, 1988, pp. 2414-2415.
- [11] S. S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence, and Robustness*, 1st ed. Englewood Cliffs, NJ: Prentice-Hall, 1989.

- [12] S. Sheikholeslam and C. A. Desoer, "Longitudinal control of a platoon of vehicles," in *Proc. Amer. Control Conf.*, vol. 1, May 1990, pp. 291-297.
- [13] —, "Design of decentralized adaptive controllers for a class of interconnected nonlinear dynamical systems," in *Proc. Conf. Decision and Control*, vol. 1, Dec. 1992, pp. 284-288.
- [14] —, "Indirect adaptive control of a class of interconnected nonlinear dynamical systems," *Int. J. Contr.*, vol. 57, no. 3, pp. 743-765, 1993.
- [15] —, "A system-level study of the longitudinal control of a platoon of vehicles," *ASME J. Dynamic Syst., Measure. Contr.*, vol. 114, no. 2, pp. 286-292, June 1992.
- [16] A. Teel, R. Kadiyala, P. Kokotovic, and S. S. Sastry, "Indirect techniques for adaptive input-output linearization of nonlinear systems," *Int. J. Contr.*, vol. 53, 1991, pp. 193-222.

Computation of Approximate Null Vectors of Sylvester and Lyapunov Operators

Ali R. Ghavimi and Alan J. Laub

Abstract—This paper describes an effective algorithm for computing approximate null vectors of certain matrix operators associated with Sylvester or Lyapunov equations. The singular value decomposition and rank-revealing QR methods are two widely used stable algorithms for numerical determination of the rank and nullity of a matrix A . These methods, however, are not readily applicable to Sylvester and Lyapunov operators since they require on the order of n^6 arithmetic operations on order n^2 data. For these problems, a variant of inverse power iteration is employed to compute orthonormal bases for singular subspaces associated with the small singular values. The method is practical since it relies only on the ability to solve a Sylvester or Lyapunov equation. Certain practical aspects are considered, and a direct refinement technique is proposed to enhance the convergence of the algorithm.

I. INTRODUCTION

The purpose of this paper is to describe an algorithm for computing approximate numerical null vectors of a matrix operator associated with the Sylvester equation $AX + XB = C$ or the symmetric Lyapunov equation $A^T X + X A^T = -Q = -Q^T$, where $X \in \mathbb{R}^{n \times m}$, $Y = Y^T \in \mathbb{R}^{n \times n}$, and the coefficient matrices are compatibly dimensioned. These equations arise in many applications in applied mathematics and control theory. For example, the solution of linear elliptic boundary value problems on rectangular domains can be written in the form of a Sylvester equation [22], [27]. The Sylvester equation also arises in connection with multi-input pole assignment problems [6], [16] and the design of reduced-order observers that achieve precise loop transfer recovery [3], [26]. The Sylvester equation also plays an essential role in many eigenproblems. Examples of such algorithms that require solution of a related Sylvester equation include invariant subspace problems, block diagonalization of a matrix [10], reordering the eigenvalues of a quasi-triangular matrix [1], and computing real square roots of a real matrix [13]. Important applications of the Lyapunov equation include stability analysis of linear systems [18], [19], model reduction

Manuscript received August 6, 1993; revised February 21, 1994. This work was supported in part by Air Force Office of Scientific Research Grant F49620-94-1-0104DEF and National Science Foundation Grant ECS-9120643.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106-9560 USA.

IEEE Log Number 9407234.

[5], [14], [21], and determination of covariance matrices in filtering and estimation.

In floating-point arithmetic with finite precision, it is quite possible that these equations are nearly singular. Equivalently, some of the small singular values of the matrix operator

$$S = I_m \otimes A + B^T \otimes I_n \in \mathbb{R}^{m \times m} \quad (1)$$

for the Sylvester equation, or

$$L = I_n \otimes A + A \otimes I_n \in \mathbb{R}^{n^2 \times n^2} \quad (2)$$

for the Lyapunov equation, are of the order of the relative machine precision (here \otimes denotes the Kronecker product [12]). In this situation, computed solutions of the matrix equations using general-purpose algorithms are usually of large norm and generally inaccurate.

Examples of ill-conditioned Sylvester and Lyapunov equations can be found in a variety of problems in control theory, especially in connection with the control of large-scale dynamical systems and problems subject to H_∞ constraints. For example, consider the deterministic optimal control problem whose quadratic performance index is of the form

$$\min_u \int_0^\infty [x^T(t)Gx(t) + u^T(t)Ru(t)]dt$$

subject to $\dot{x}(t) = Ax(t) + Bu(t)$ with $x(0) = x_0$, where $R = R^T > 0$ and $G = G^T \geq 0$ are symmetric matrices, B is an $n \times m$ input matrix, and $u(t)$ is the control law. Under the usual stabilizability and detectability assumptions for (A, B) and (A, G) , respectively, the optimal feedback control for this cost function is given by $u_{\text{opt}}(t) = -R^{-1}B^T Px(t)$ and the associated minimum cost is $J_{\text{min}} = x_0^T Px_0$, where P is the unique symmetric nonnegative definite stabilizing solution of the algebraic Riccati equation

$$PA + A^T P - PBR^{-1}B^T P + G = 0. \quad (3)$$

A reliable algorithm [20] for solving (3) involves a Schur decomposition of the related Hamiltonian matrix

$$\begin{bmatrix} A & -BR^{-1}B^T \\ -G & -A^T \end{bmatrix} = QSQ^T; \quad S = \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{bmatrix}$$

where Q is orthogonal and S is quasi-upper triangular. The eigenvalues of S are ordered via additional orthogonal similarities if necessary so that S_{11} and S_{22} are quasi-upper triangular and S_{11} contains the stable eigenvalues. Reordering of a Schur form requires solutions of certain Sylvester equations where the coefficient matrices are related to the diagonal blocks of S . For the sake of illustration, let

$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ 0 & \hat{A}_{22} \end{bmatrix}$$

where $\hat{A}_{11} \in \mathbb{R}^{p \times p}$ and $\hat{A}_{22} \in \mathbb{R}^{q \times q}$, with $p, q = 1$, or 2 . The problem is to find an orthogonal matrix $N \in \mathbb{R}^{(p+q) \times (p+q)}$ such that

$$\hat{A} = \begin{bmatrix} \hat{A}_{22} & \hat{A}_{12} \\ 0 & \hat{A}_{11} \end{bmatrix} = N^T \hat{A} N$$

and \hat{A}_{11} and \hat{A}_{22} have the same eigenvalues. In other words, \hat{A} is similar to \hat{A} and the eigenvalues of the $(1, 1)$ and $(2, 2)$ subblocks have been interchanged. It can be shown that N is obtained from the QR factorization

$$\begin{bmatrix} -X \\ I_q \end{bmatrix} = N \begin{bmatrix} U \\ 0 \end{bmatrix}$$

where $X \in \mathbb{R}^{p \times q}$ solves the Sylvester equation $\hat{A}_{11}X - X\hat{A}_{22} = \hat{A}_{12}$. This Sylvester equation can be nearly singular for a number of reasons. A case of particular interest is when \hat{A}_{11} and \hat{A}_{22} contain

eigenvalues that are reflections of each other with respect to the imaginary axis and are close to the imaginary axis. That is, if $\alpha \pm \beta j$ are the eigenvalues of \hat{A}_{11} , then $-\alpha \pm \beta j$ are the eigenvalues of \hat{A}_{22} and $0 < \alpha \ll 1$. In this case, the associated Sylvester equation is ill conditioned, N can be computed only approximately, and the zero block in \hat{A} is actually nonzero. As a result, the computed solution of (3) may not have sufficient accuracy and, consequently, may induce instability in the computed closed-loop matrix.

Computed solutions of (3) may be refined via Newton's method [17]. This involves rewriting (3) as $P(A - FP) + (A - FP)^T P = -G - PFP$, where $F = BR^{-1}B^T$, and suggests the iteration

$$P_{k+1}(A - FP_k) + (A - FP_k)^T P_{k+1} = -G - P_k F P_k \quad (4)$$

which can be shown to be equivalent to Newton's method for solving (3). It can further be shown that if $A - FP_0$ is stable, the sequence generated by this iteration ensures the stability of the "closed-loop" matrices $A - FP_k$ at each step and $\lim_{k \rightarrow \infty} P_k = P$. Potential difficulty may arise, however, when the Lyapunov equations in (4) are ill conditioned. This occurs if, for example, $A - FP_k$ has an eigenvalue very close to the imaginary axis. In this case, the procedure can give inaccurate solutions which may cause the process to leave the region of convergence for Newton's method.

An alternative approach for solving ill-conditioned Sylvester and Lyapunov equations is to consider least-squares-type solutions. This then allows a computed solution to be expressed more accurately in terms of a part that is purged of the numerical or approximate null vectors (in a sense to be defined below) and a part that is spanned by the numerical null vectors. As in the case of linear systems, however, a knowledge of approximate null vectors is needed as a basic tool for any least-squares solver.

The numerical rank and nullity of a matrix are best understood in terms of its singular value decomposition (SVD). Many theorems and results in linear algebra depend explicitly or implicitly on knowing the rank of a matrix. Computational difficulties may arise, however, when working in finite precision. For example, a matrix that is known to be nonsingular may be nearly singular. It is therefore appropriate to define numerical ϵ -rank and ϵ -nullity for some small ϵ . In [8] and [10], these concepts are defined in terms of the SVD. Specifically, it is shown that the ϵ -nullity of a matrix is equal to the number of singular values that are less than a prespecified ϵ in magnitude. Moreover, the numerical null vectors are best characterized in terms of the associated singular vectors. In this respect, the remainder of the paper adopts this convention and we refer to ϵ -null vectors of a nearly singular matrix as the approximate or numerical null vectors of that matrix.

The organization of the paper is as follows. Section II describes a variant of the simultaneous iteration method for the dominant eigenvectors of a Hermitian matrix [23] to implement an algorithm for computing approximate null vectors of a matrix. A simple refinement technique is then proposed to enhance convergence. Section III specializes the method to nearly singular Sylvester and Lyapunov equations. Certain practical aspects also are considered to exploit efficiency. Section IV presents a numerical example to demonstrate further the effectiveness of the algorithm and concluding remarks are given in Section V.

II. SIMULTANEOUS INVERSE ITERATION METHOD

This section considers the subspace version of an inverse iteration algorithm for finding the smallest singular value of a matrix [28]. The method is analogous to the orthogonal iteration technique (a generalization of the power method) for finding a prescribed set of the dominant eigenvalues of a Hermitian matrix [23]. The differences can be resolved by first viewing the singular values of a matrix A in

ms of the eigenvalues of the Hermitian matrix $A^T A$, and secondly applying the orthogonal iteration to A^{-1} rather than A . Note that similar ideas can also be found in [15], [28] and other scattered work throughout the literature. An expository outline of the main ideas is presented here for the sake of completeness.

Suppose that $A \in \mathbb{R}^{n \times n}$ is nearly rank m -deficient. That is, A has small singular values that are of the order of the relative machine precision. Define an SVD of A by

$$A = U \Sigma V^T = [U_1 \quad U_2] \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} [V_1 \quad V_2]^T \quad (5)$$

where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{n-m} \gg \sigma_{n-m+1} \geq \dots \geq \sigma_n > 0$ and the partitions are conformal. It then follows that the orthonormal columns of U_2 and V_2 are bases for the approximate left and right null vectors of A , respectively [10]. When the SVD of A is not available explicitly, the near-singular subspaces U_2 and V_2 can be approximated implicitly by U and V , respectively as follows.

Simultaneous Inverse Iteration Algorithm

- Choose $U_0 \in \mathbb{R}^{n \times m}$ with orthonormal columns (almost arbitrary)
- Loop until convergence
 - 1) Solve $U_{i+1} = U$
 - 2) $V_{i+1} = V_{i+1} T_{i+1}$ (QL factorization)
 - 3) Solve $U_{i+1}^T U_{i+1} = I$
 - 4) $U_{i+1} = U_{i+1} R_{i+1}$ (QL factorization)
- End loop
- $U \leftarrow$ converged U_{i+1} and $V \leftarrow$ converged V_{i+1}

The QL factorization in the above algorithm is a natural extension of the well known QR factorization. This decomposition is of the form $U = QI$ where $Q \in \mathbb{R}^{n \times m}$ has orthonormal columns and $I \in \mathbb{R}^{m \times m}$ is lower triangular. It is useful however to focus attention further on the unique choice of the QL factorization of $U \in \mathbb{R}^{n \times m}$ (or possibly square) by requiring $I \in \mathbb{R}^{m \times m}$ to be lower triangular with positive diagonal elements. The uniqueness of such a decomposition follows by similar analysis for the "skinny" QR factorization [10].

The convergence of the simultaneous inverse iteration algorithm parallels similar arguments for the convergence of the QR method in [28] or orthogonal iteration in [23]. A detailed convergence proof of the algorithm is also given in [9]. Several of the key points are summarized as follows. The above algorithm converges to a set of m singular vectors of A almost independent of the choice of U_0 . The convergence might be unstable however, in the sense that some of the converged singular vectors do not correspond to those associated with the small singular values. This happens when $\mathcal{R}(U_0)$ is orthogonal to at least one of the singular vectors in U_2 . In practice, this situation is somewhat unlikely to occur since rounding errors usually turn unstable convergence into a delayed stable one. In any case, a countermeasure should be taken to overcome this undesirable effect. It is easy to verify that the last column of U_{i+1} (i.e., Step 4 of the above loop) is most accurate in the sense that it has the most contributions from the singular vectors associated with the small singular values. On the other hand, the first column of U_{i+1} is the least accurate since it contains the components of singular vectors corresponding to singular values larger than σ_{n-m+1} . This suggests that the first column of U_{i+1} can be replaced by a random vector that is orthogonal to the last $m-1$ columns. This process can be terminated once the convergence of the first column vector is ascertained.

If the singular values satisfy the condition stated after (5), the columns of U_k and V_k converge to vectors that are linear combina-

tions of the left and right singular vectors, respectively. In the span of the same subspaces. Furthermore, when the small singular values are all distinct and well separated from each other, then the limiting values of U_k and V_k converge to the individual singular vectors denoted by the columns of U_2 and V_2 , respectively. The convergence rate in this case is quadratic with a ratio of $\max \left\{ \frac{\sigma_{n-m+1}}{\sigma_{n-m}}, \frac{\sigma_{n-m}}{\sigma_{n-m+1}} \right\}$ with $i > j$. Even if the small singular values are poorly separated, U_k and V_k converge to a linear span of the singular vectors in U_2 and V_2 at a rate of $(\sigma_{n-m+1}/\sigma_{n-m})^2$, a rapid convergence rate by assumption. Note that for singular values of nearly equal magnitude additional iterations for convergence to the individual singular vectors is generally slow and computationally not tractable. In other words, while the individual singular vectors associated with a set of clustered singular values are poorly determined, the singular subspaces spanned by them are well determined. Moreover, a direct refinement procedure may also be utilized to both enhance the convergence and reduce the computational cost.

The following refinement scheme is somewhat similar to a procedure used by Stewart for accelerating the convergence of the approximate dominant eigenvectors of a matrix via the orthogonal iteration method [23]. Similar techniques have also been used in connection with finding complex conjugate eigenvalues of a real matrix [28], diagonalization of a real symmetric matrix [10], and calculation of the dominant invariant subspace of a nonsymmetric matrix [2]. These methods are generally known as Ritz acceleration techniques.

Suppose that the orthonormal columns of U and V are approximations to the left and right singular vectors, respectively, corresponding to the set of small singular values of A whose SVD is given by (5). Define $P, Q \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times m}$ by

$$P = U^T U = \begin{bmatrix} U_1^T U \\ U_2^T U \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$$

and

$$Q = V^T V = \begin{bmatrix} V_1^T V \\ V_2^T V \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \quad (6)$$

$$B = U^T A V = U^T U \Sigma V^T V = P^T \Sigma Q = P_1^T \Sigma_{11} Q_1 + P_2^T \Sigma_{22} Q_2 \quad (7)$$

Moreover, denote the SVD of B by $B = M \Lambda \Lambda^T$ where M and Λ are orthogonal and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$ whose entries satisfy $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$. When $P_1 = Q_1 = 0$, P_2 and Q_2 are orthogonal. Thus $\Sigma_{22} = \Lambda$. Furthermore, if the singular values are distinct, then corresponding singular vectors can be recovered by choosing $M = P_2^T$ and $\Lambda = Q_2^T$ since $U M = U P M = U P_2 P_2^T = U_2$. Similarly, $V \Lambda = V$. The above suggests that when the column spaces of U and V are good approximations of the spans of the singular subspaces associated with the small singular values, P_1 and Q_1 are nearly zero, P_2 and Q_2 are nearly orthogonal and $\Sigma_{22} \approx \Lambda$. In addition, $U M$ and $V \Lambda$ are better approximations to U_2 and V_2 than U and V , respectively. The following states the algorithm in a formal manner.

A Refinement Method for Approximate Singular Vectors

- Start with U and V as approximations to U_2 and V_2 in (5) respectively
- Form $B = U^T A V$
- Compute the SVD of $B = M \Lambda \Lambda^T$
- $U \leftarrow U M$ and $V \leftarrow V \Lambda$

Details of the above procedure parallel similar results in [23] and has been examined closely in [9]. Thus, only certain highlights of the method are summarized here. Let $\Delta \Sigma = \Lambda - \Sigma_{22}$ denote the deviation of the refined singular values from the actual ones.

Then $|\Delta\Sigma| = |\Lambda - \Sigma_{22}| \leq \sigma_1 \|P_1\| \|Q_1\| + \sigma_{n-m+1} (\|P_1\|^2 + \|P_1\|^2 \|Q_1\|^2 + \|Q_1\|^2)$, where the absolute value of a matrix is in the componentwise sense. This bound implies that for small values of P_1 and Q_1 , the refined singular values are therefore good approximations of their actual values. As far as the accuracy of the refined singular vectors is concerned, it is seen that $U(V)$ and $UM(VN)$ span the same space. Therefore, the refined singular vectors are just as good approximations of the singular subspaces of interest as U and V are. Furthermore, UM and VN are better enhancements of the individual singular vectors than U and V , respectively. This can be seen from $(UM)^T A(VN) = M^T U^T A V N = M^T B N = \Lambda \approx \Sigma_{22}$, whereas $B = U^T A V$ is generally a dense matrix. Moreover, the refined singular vectors tend to converge at a faster rate than the original ones. This suggests that the refinement algorithm can be employed after a few steps of simultaneous iteration to further enhance the convergence rate of the iterates. The next section incorporates the above results to construct bases for the approximate null vectors of nearly singular Sylvester and Lyapunov equations.

III. INVERSE ITERATION FOR SYLVESTER AND LYAPUNOV EQUATIONS

This section extends the simultaneous inverse iteration method to compute orthonormal bases for approximate null spaces of matrix operators associated with nearly singular Sylvester and Lyapunov equations. The extension follows by viewing the Sylvester or Lyapunov equation as the linear system $Sx = c$ or $Ly = q$, respectively, where S and L are as specified in (1) and (2), and $x = \text{vec}(X)$, $c = \text{vec}(C)$, $y = \text{vec}(Y)$, and $q = \text{vec}(Q)$. Note that vec of an $m \times n$ matrix M is the mn -vector formed by successively stacking the n columns of M on top of each other. Since the Lyapunov equation is a special case of the Sylvester equation, the results are presented for the Sylvester equation. The method applies to the Lyapunov equation with only a slight and obvious modification.

Suppose that $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{m \times m}$ are given and that the associated matrix operator S in (1) is nearly singular. Algorithms for deflated solutions of nearly singular linear systems are given in [7], [25]. Note, however, that while the extension of the simultaneous inverse iteration method to a nearly singular Sylvester equation seems natural, certain steps are crucial. Potential difficulties in using the method derive from the fact that numerical nullities are not known *a priori* and that the matrix S in (1) generally should not be formed explicitly. It is therefore logical to construct bases for the numerical null spaces one vector at a time. Successive formation of the approximate null vectors involves normalizations with respect to the Frobenius norm, and orthogonalizations with respect to the inner product $\langle w, z \rangle = w^T z = \text{tr}(W^T Z)$, where $w = \text{vec}(W)$ and $z = \text{vec}(Z)$.

Let the orthonormal columns of $P \in \mathbb{R}^{m \times n}$ and $Q \in \mathbb{R}^{n \times m}$ denote the left and right singular vectors of S , respectively, associated with the ν small singular values. The following algorithm constructs approximations to P and Q by U and V , respectively.

Simultaneous Inverse Iteration for Nearly Singular Sylvester Equations

- For nullity $\nu \geq 1$, choose $U_0 \in \mathbb{R}^{m \times \nu}$ with orthonormal columns.
- Loop until convergence
 - 1) Solve $S\tilde{V} = U_0$ for \tilde{V}
 - 2) $\tilde{V} = \tilde{V}T$ (QL factorization)
 - 3) Solve $S^T \tilde{U} = \tilde{V}$ for \tilde{U}
 - 4) $\tilde{U} = \tilde{U}R$ (QL factorization)
 - 5) $G = \tilde{U}^T S \tilde{V}$
 - 6) $G = M \Lambda N^T$ (SVD)

- 7) $U \leftarrow UM$ and $V \leftarrow VN$
- 8) If $\lambda_j > \text{tol}$, a user-specified tolerance, for $1 \leq j \leq \nu$,
 - a) $U \leftarrow U(:, j+1 : \nu)$ (truncate the first j columns)
 - b) $V \leftarrow V(:, j+1 : \nu)$ (truncate the first j columns)
 - c) $\nu = j$
 - d) Stop
 - Else
 - e) $\nu = \nu + 1$
 - f) $U_0 = [\tilde{U}^T U]$ such that $U_0^T U_0 = I_\nu$
 - g) Return to Innerstep 1).

The following provides certain key details of the actual implementation of the algorithm. As was mentioned above, the large structured matrix S is generally not available explicitly. Therefore, appropriate manipulations are necessary for operations involving S . Let x_k and y_k be the corresponding left and right singular vectors of S associated with the singular value σ_k . That is, $Sy_k = \sigma_k x_k$ and $S^T x_k = \sigma_k y_k$. It then follows that $A Y_k + Y_k B = \sigma_k X_k$ and $A^T X_k + X_k B^T = \sigma_k Y_k$, where $X_k = \text{unvec}(x_k) \in \mathbb{R}^{n \times m}$, $Y_k = \text{unvec}(y_k) \in \mathbb{R}^{m \times m}$, and $\text{unvec}(\cdot)$ is the operator that "undoes" the vec operator, i.e., forms a matrix from a column vector stack. Furthermore, Innerstep 1) is a linear system with multiple right-hand sides. The equation is of the form $SW = Z$, where $W = (w_1, \dots, w_\nu)$ and $Z = (z_1, \dots, z_\nu) \in \mathbb{R}^{m \times \nu}$. The k th column of W is found by solving $Sw_k = z_k$. This can be written as $A W_k + W_k B = Z_k$, where $W_k = \text{unvec}(w_k) \in \mathbb{R}^{n \times m}$ and $Z_k = \text{unvec}(z_k) \in \mathbb{R}^{m \times m}$. Thus, the columns of W are obtained by successively solving the associated Sylvester equations and setting $w_k = \text{vec}(W_k)$. Moreover, Innerstep 3) is also a linear system of the form $S^T W = Z$. Hence, W can be solved by rewriting the Sylvester equation of the above loop as $A^T W_k + W_k B^T = Z_k$. If the Bartels–Stewart algorithm [4] or the Hessenberg–Schur method [11] is used, these equations can be solved fairly inexpensively with the aid of previously computed factorizations of A and B . Innerstep 5) is of the form $G = (g_{ij}) = W^T S Z \in \mathbb{R}^{\nu \times \nu}$. Let e_i denote the k th elementary unit vector. Then the entries of G are given by

$$g_{ij} = e_i^T W^T S Z e_j = w_i^T S z_j = (\text{vec}(W_i))^T \text{vec}(AZ_j + Z_j B) \\ = \text{tr}(W_i^T (AZ_j + Z_j B)).$$

Innersteps 8a) and 8b) are MATLAB notations for overwriting U and V by deleting their first j columns. Innerstep 8c) indicates that the numerical rank of S is ν . Innerstep 8f) is realized by letting $w = z - U U^T z$, where $z \in \mathbb{R}^{m \times m}$ is a random vector. Then $\bar{w} = \frac{w}{\|w\|}$ and $\bar{U} = \text{unvec}(\bar{w})$.

IV. NUMERICAL RESULTS

The following example is presented to illustrate the effectiveness of the results. All of the computations were performed using MATLAB implemented on a Sparc 2 workstation. The relative machine precision reported by MATLAB is about 2.22×10^{-16} .

Let A and B be of the form $A = W \Lambda W^T$ and $B = Z \Gamma Z^T$, where $W \in \mathbb{R}^{n \times n}$ and $Z \in \mathbb{R}^{m \times m}$ are random orthogonal matrices. Moreover, $\Lambda = \text{diag}(n, n-1, \dots, 3, \alpha_2, \alpha_1)$ and $\Gamma = \text{diag}(m, m-1, \dots, 3, \beta_2, \beta_1)$, where $\alpha_1, \alpha_2, \beta_1$, and β_2 are between 3.0×10^{-14} and 1.0×10^{-15} . Note that W and Z in this example are formed via the QR factorization of random matrices. A thorough discussion of computing random orthogonal matrices can be found in [24]. Since W and Z are orthogonal, the eigenvalues of A and B are given by the diagonal entries of Λ and Γ , respectively. Furthermore, $S = M D M^T$ where $M = (Z^T W)$ is orthogonal and $D = I_m - \Lambda + \Gamma - I_n$ is diagonal. Hence, the eigenvalues of S are given by the sum of any two diagonal elements of Λ and Γ . Moreover, the singular

values of S are the absolute values of its eigenvalues, i.e., S is normal. This implies that S has four small singular values, namely, $(\alpha_1 + \beta_1)$, $(\alpha_1 + \beta_2)$, $(\alpha_2 + \beta_1)$, and $(\alpha_2 + \beta_2)$ for any $m, n \geq 3$. Thus the Sylvester equation $AX + XB = C$ is nearly singular. For the sake of illustration, the dimension m was kept constant at 10 while n was varied from 3 to 100 in increments of one, and vice versa. Approximate null vectors of S were found directly by computing an SVD of S and selecting the singular vectors associated with the small singular values. Next, the proposed inverse iteration algorithm was used to compute the singular subspaces of interest. In each case, only one or two iterations at most was necessary to achieve convergence. The results were compared numerically by measuring the angles between subspaces [10] obtained directly (via SVD) and those by the inverse iteration algorithm. It was observed that null vectors obtained by the two methods were essentially the same and were good approximations of each other. The computational advantage of the inverse iteration algorithm over the SVD method was apparent as the dimensions m and n increased. This results from the fact that the direct method requires the SVD of an $nm \times nm$ matrix, whereas the inverse iteration algorithm requires solving only a few $n \times m$ Sylvester equations. In terms of floating-point operations, the SVD method and the inverse iteration algorithm therefore require on the order of $\alpha m^3 n^3$ flops and $\beta(m^3 + n^3) + \gamma(m^2 n + mn^2)$, respectively, where α , β , and γ are some modest constants. For example, if $m = n = 10$, the number of flops needed by the SVD method is on the order of 10^6 compared to 10^3 for the inverse iteration algorithm. Even greater efficiencies can be realized when the computed factorizations of A and B in the first step are used in succeeding steps.

V. CONCLUDING REMARKS

An analysis has been given of computation of approximate numerical null vectors for matrix operators associated with Sylvester equations. The method is reliable, efficient, and requires nothing but the ability to solve linear matrix equations. Computational complexity can be reduced by saving computed decompositions of the coefficient matrices. The method becomes relatively more expensive, however, as the number of small singular values increases.

One advantage of the inverse iteration algorithm lies in its modularity. Note that the linear systems in the innersteps 1) and 3) of the algorithm are independent processes. This means that the method can be easily extended to the Lyapunov equation $AX + XA^T = Q$ or the discrete-time Sylvester equation $AXB + X = C$ and the discrete-time Lyapunov equation $AXA^T - X = Q$. In fact, it can be applied to any matrix equation that can be formulated as a standard linear system. Another important factor is that the choice of linear solvers is also an independent step in the algorithm. This is significant since specialized solvers can be employed in place of conventional ones when appropriate.

REFERENCES

- [1] Z. Bai and J. W. Demmel, "On swapping diagonal blocks in real Schur form," *Lin. Alg. Appl.*, vol. 186, pp. 73-96, 1993.
- [2] Z. Bai and G. W. Stewart, "SPRIT—A Fortran subroutine to calculate the dominant invariant subspace of a nonsymmetric matrix," CS TR-

- 2908, Dept. of Computer Science, Univ. Maryland, UMIACS, Tech. Rep. TR-92-61, May 1992.
- [3] J. B. Barlow, M. M. Monahemi and D. P. O'Leary, "Constrained matrix Sylvester equations," *SIAM J. Matrix Anal. Appl.*, vol. 13, pp. 1-9, 1992.
- [4] R. H. Bartels and G. W. Stewart, "Algorithm 432: Solution of the matrix equation $AX + XB = C$," *Commun. ACM*, vol. 15, pp. 820-826, 1972.
- [5] D. S. Bernstein and D. C. Hyland, "The optimal projection equations for reduced-order state estimation," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 583-585, 1985.
- [6] S. P. Bhattacharyya and E. DeSouza, "Pole assignment via Sylvester's equation," *Sys. Contr. Letts.*, vol. 1, pp. 261-263, 1982.
- [7] T. F. Chan, "Deflated decomposition of solutions of nearly singular systems," *SIAM J. Numer. Anal.*, vol. 21, pp. 738-754, 1984.
- [8] L. V. Foster, "Rank and null space calculations using matrix decomposition without column interchanges," *Lin. Alg. Appl.*, vol. 74, pp. 47-71, 1986.
- [9] A. R. Ghavimi, "Iterative methods for large-scale and nearly singular matrix equations in control theory," Ph.D. thesis, Univ. of California, ECE Dept., Santa Barbara, CA, June 1993.
- [10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd. ed. Baltimore, MD: Johns Hopkins Univ. Press 1989.
- [11] G. H. Golub, S. Nash, and C. F. Van Loan, "A Hessenberg-Schur method for the problem $AX + XB = C$," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 909-913, 1979.
- [12] A. Graham, *Kronecker Products and Matrix Calculus with Applications*. New York: Wiley, 1981.
- [13] N. J. Higham, "Computing real square roots of a real matrix," *Lin. Alg. Appl.*, vol. 88/89, pp. 405-430, 1987.
- [14] D. C. Hyland and D. S. Bernstein, "The optimal projection equations for fixed-order dynamic compensation," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 1034-1037, 1984.
- [15] A. Jennings, *Matrix Computation for Engineers and Scientists*. New York: Wiley, 1977.
- [16] L. H. Keel, J. A. Fleming, and S. P. Bhattacharyya, "Minimum norm pole assignment via Sylvester's equation," in *Linear Algebra and Its Role in Systems Theory*, (Contemporary Math. Series.) vol. 47. Providence, RI: Amer. Math. Soc. 1985, pp. 265-272.
- [17] C. S. Kenney, A. J. Laub, and M. Wette, "Error bounds for Newton refinement of solutions to algebraic Riccati equations," *Math. Contr. Sig. Sys.*, vol. 3, pp. 211-224, 1990.
- [18] D. L. Kleinman, "An easy way to stabilize a linear constant system," *IEEE Trans. Automat. Contr.*, vol. AC-15, p. 692, 1970.
- [19] J. La Salle and S. Lefschetz, *Stability by Liapunov's Direct Method*. New York: Academic, 1961.
- [20] A. J. Laub, "A Schur method for solving algebraic Riccati equations," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 913-921, 1979.
- [21] B. C. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 17-31, 1981.
- [22] G. Starke and W. Niethammer, "SOR for $AX - XB = C$," *Lin. Alg. Appl.*, vol. 154, pp. 355-375, 1991.
- [23] G. W. Stewart, "Accelerating the orthogonal iteration for the eigenvectors of a Hermitian matrix," *Numer. Math.*, vol. 13, pp. 362-376, 1969.
- [24] —, "The efficient generation of random orthogonal matrices with an application to condition estimators," *SIAM J. Numer. Anal.*, vol. 17, pp. 403-409, 1980.
- [25] —, "On the implicit deflation of nearly singular systems of linear equations," *SIAM J. Sci. Stat. Comp.*, vol. 2, pp. 136-140, 1981.
- [26] C. C. Tsui, "A new approach to robust observer design," *Int. J. Control*, vol. 47, pp. 745-751, 1988.
- [27] F. Wan, "An in-core finite difference method for separable boundary value problems on a rectangle," *Studies in Appl. Math.*, vol. 52, pp. 103-113, 1973.
- [28] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*. Oxford: Oxford Univ. Press, 1965.



THE 34TH IEEE CONFERENCE ON DECISION AND CONTROL

New Orleans Hilton Riverside and Towers
New Orleans, Louisiana
December 13-15, 1995



The IEEE CSS Conference on Decision and Control (CDC) is the annual meeting of the IEEE Control Systems Society (CSS), conducted in cooperation with the Society for Industrial and Applied Mathematics (SIAM) and the Operations Research Society of America (ORSA). The thirty-fourth CDC will be held December 13-15, 1995, with tutorial workshops preceding the conference on Monday and Tuesday, December 11-12. The Conference General Chairman is Panos Antsaklis of the University of Notre Dame, and the Program Chairman is Edward Kamen of the Georgia Institute of Technology. The venue is the New Orleans Hilton Riverside and Towers, located on the Mississippi River immediately west of Canal Street. A short walk through a landscaped Riverwalk brings one to the French Quarter, scene of New Orleans nightlife and musical history.

CALL FOR CONTRIBUTED PAPERS AND INVITED SESSIONS

The IEEE CDC will include both contributed and invited sessions and a full Proceedings will be published. Contributed papers are hereby solicited in all aspects of the theory and applications of systems, including decision-making, control, adaptation, optimization, industrial automation, and manufacturing. Invited sessions are also solicited in new developments in these and related areas. All submissions are due 1 MARCH 1995.

CONTRIBUTED PAPERS:

The Program Committee is soliciting both regular and short contributed papers for presentation at the conference and publication in the Proceedings. All submissions must follow the guidelines below.

REGULAR PAPERS. Regular papers describe completed work in some detail. Authors should submit five (5) copies of the full paper for review to the Conference Publications Editor at the address below.

SHORT PAPERS. Short papers describe important recent or preliminary results which require limited length for their development. Authors should submit four (4) copies of a 4-6 page detailed summary, including references, for review to the Program Vice-Chair for Short Papers at the address below.

INVITED SESSIONS:

The Program Committee is soliciting proposals for invited sessions. Cohesive sessions focusing on new or emerging topics in the above-listed areas are particularly encouraged, and will have priority over those of a classical or mainstream flavor. Session organizers should submit four (4) copies of proposals for invited sessions for review to the Program Vice-Chair for Invited Sessions at the address below. Organizers should contact the Vice-Chair by 1 FEBRUARY 1995 stating their intent to submit a proposal.

Proposals for invited sessions should contain the names, affiliations, and complete mailing addresses of the session organizer(s), chairperson, co-chairperson, and all authors (with corresponding authors identified) along with a list of paper titles. The organizers must include in the proposal: (1) a clear statement of the topic and purpose of the session and (2) a description of how the papers form a cohesive, well-integrated exploration of the session topic. Detailed extended summaries, following the submission guidelines below and consisting of a minimum of 2000 words, covering all contributions of the paper in sufficient depth to permit an informed review, must be included for each paper. The initial paper may be a tutorial or survey which can be allotted twice the usual time for presentation. The usual number of papers in a session is six. Organizers will be contacted near 1 June 1995 concerning the tentative disposition of their session.

SUBMISSION GUIDELINES AND ACCEPTANCE INFORMATION

Each submitted regular paper, short paper extended summary, or invited session proposal must be headed with a title, the names, affiliations and complete mailing addresses of all authors (or organizers for invited sessions), a list of three keywords, and the statement "34th CDC". The first named author of each paper will be used for all correspondence unless otherwise requested, if possible an electronic mail address and fax number of the corresponding author should be included. Incomplete submissions or submissions by fax or electronic mail cannot be accepted.

Final selection of papers and invited sessions will be announced in mid-July 1995. Authors of all accepted papers will be provided with instructions for preparation of manuscripts for the Proceedings. Authors should limit their manuscripts to six Proceedings pages (approx. 6000 words) for regular papers or two Proceedings pages (approx. 2000 words) for short papers. There will be a mandatory page charge for each additional page. Authors of accepted papers are expected to attend the CDC to present their work.

ADDRESSES FOR SUBMISSIONS

Regular Papers:

Prof. M. Peshkin, 34th CDC
Editor, Conference Publications
Mechanical Engr. Dept.
Northwestern University
2145 Sheridan Road
Evanston, Ill 60208-3111
Tel: (708) 467-2666
Fax: (708) 491-3915
email: cdc@nwu.edu

Short Papers:

Prof. J. Jim Zhu, 34th CDC
VC for Short Papers
3221 CEBA Bldg.
Rem. Sens./Image Proc. Lab.
Louisiana State University
Baton Rouge, LA 70803
Tel: (504) 388-6826
email: zhu@sun-ra.rsp.lsu.edu

Invited Sessions:

Prof. W. J. Rugh, 34th CDC
VC for Invited Sessions
Elect. & Computer Engr. Dept.
Johns Hopkins University
Baltimore, MD 21218
Tel: (410) 516-7004
email: wjr@jhuunix.hcf.jhu.edu

General Information:

Panos J. Antsaklis, 34th CDC
General Chair
Dept. of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
Tel: (219) 631-5792
Fax: (219) 631-4393
Panos.J.Antsaklis@nd.edu

SCHEDULE SUMMARY:

1 Feb. 1995	Deadline for statement of intent to submit invited session proposals
1 March 1995	Deadline for submission of contributed papers and invited session proposals
1 June 1995	Tentative notification of invited session organizers
Mid-July 1995	Notification regarding acceptance of papers and invited sessions
Early Aug. 1995	Instructions for manuscript preparation sent to authors
Mid-Sept. 1995	Camera-ready papers due at the printer

Scanning the Issue*

Observer Design for Nonlinear Systems with Discrete-Time Measurements, *Moraal and Grizzle*.

Observer Design for Nonlinear Systems has been a hot topic in nonlinear control theory since the early 1980's when Krener and Isidori published their seminal paper of linearization of nonlinear systems by output injection and the relevance of this concept to nonlinear observer design. Much effort has been devoted to finding a nonlinear analog of the now standard observer construction procedures in linear systems theory. Predominantly, the method of attack has been to try out some nonlinear generalizations of linear techniques. Even though there have been sporadic successes, these attempts have not led to a rich theory.

The work of Moraal and Grizzle is aimed at moving away from this approach and focussing more on techniques that have to do with inverting nonlinear maps. The notion of uniform γ observability essentially says that there are a fixed number of inputs and a system of equations the solution of which yields the initial state of the system uniquely. Now the problem of solving the equation can be approached as any numerical analyst would do by resorting to numerical map inversion techniques such as the Newton-Raphson method. A suitable generalization leads to a system which the authors call a Newton Observer. This procedure is described in great detail in the paper and mild conditions that guarantee convergence are given. It is shown that Newton Observers are closely related to Kalman filters. An example to illustrate the concept is given in the form of an application to a bioreactor system.

A Probabilistic Approach to Multivariable Robust Filtering and Open-Loop Control, *Ohlin Ahlen and Sternad*.

A new approach to robust filtering, prediction and smoothing of discrete time signal output of modeling errors is presented. A simple design based on the minimization of the squared estimation error averaged both with respect to model errors and noise is obtained. An optimal robust filter is calculated by means of an averaged spectral factorization and an unilateral diophantine equation.

The robust estimator is referred to as a cautious Wiener filter. It turns out to be only slightly more complicated to design than an ordinary Wiener filter. The methodology can be applied to any open loop filtering or control problem. This paper illustrates this for the design of robust multivariable feedforward regulators, decoupling, and model matching filters.

A New Model for Control of Systems with Friction, *anudas de Wit, Olsson Astrom, and Lischinsky*.

The modeling of friction is an important concern in the design of control systems for high quality servo mechanisms. Friction can lead to tracking errors, limit cycles and undesired stick-slip motion. A high gain control loop can sometimes be used to compensate for friction, but in other circumstances it is necessary to compensate for the effects of friction based on a suitable friction model. Friction models based in static maps, such as Coulomb and viscous friction, may not be satisfactory for applications such as high precision pointing or low velocity tracking. Dynamic friction models are

generally better for such applications and a number of such models have been developed. In this paper the authors introduce a new dynamic model for friction which is consistent with experimentally observed effects. The model properties relevant to control systems are investigated and discussed. The authors show that the model predicts the limit cycle behavior sometimes observed in servos with PID controls. They also use the model in the design of a friction observer to compensate for friction in position and velocity tracking control.

Adaptive Nonlinear Design with Controller-Identifier Separation and Swapping, *Krstic and Kokotovic*.

This paper introduces a modular approach to adaptive control design for a class of nonlinear systems in parametric-strict feedback form. By achieving a separation between the designs of the adaptive controller and the parameter identifier, it is shown that the stability properties of the certainty equivalence laws can be enhanced thereby enabling a wide variety of adaptive update laws to tune the control. The former is designed such that the states of the nonadaptive system are bounded if the parameter errors and their derivatives are bounded, the latter guarantees the boundedness of the parameter errors and their derivatives. A nonlinear version of the swapping lemma is used to facilitate the proofs.

Approximate Decoupling and Asymptotic Tracking for MIMO Systems, *Godbole and Sastry*.

This paper proposes an algorithm for approximate input-output decoupling of nonlinear MIMO systems where the standard decoupling algorithm produces poor results. The systems considered are regular, but are numerically ill posed in that application of the exact decoupling algorithm requires inversion of an ill conditioned matrix. The algorithm is numerically robust and also avoids pole-zero cancellations thus providing controllers with reasonable gain even when the system has far off zeros. The general algorithm here is motivated by the work of various authors on physical examples such as flight control, reaction kinetics, and electric motors.

A Generalized Orthonormal Basis for Linear Dynamical Systems, *Heuberger Van den Hof, and Bosgra*.

In many areas of signal, system, and control theory, orthogonal functions play an important role in analysis and design. In this paper, it is shown that there exist orthogonal functions that are generated by stable linear dynamical systems and that constitute an orthonormal basis for the signal space l_2 . The orthogonal functions can be considered as generalizations of, e.g., the pulse functions, Laguerre functions, and Kautz functions, and give rise to an alternative series expansion of rational transfer functions. It is shown how these generalized basis functions can be used to increase the speed of convergence in a series expansion, i.e., to obtain a good approximation by retaining only a finite number of expansion coefficients. Consequences for identification of expansion coefficients are analyzed, and a bound is formulated on the error that is made when approximating a system by a finite number of expansion coefficients.

*This section is written by the Transactions Editorial Board.

H_∞ Control Via Measurement Feedback for General Nonlinear Systems, Isidori and Kang.

The H_∞ control problem is a well-motivated problem for a wide variety of engineering applications. The problem is very well studied for linear systems. A comprehensive and elegant solution is available for linear systems and is well documented in various recent papers. The generalization of these results to nonlinear systems is an

important problem. The authors provide a solution to the nonlinear version of the H_∞ control problem with measurement feedback (i.e., the partial observations case), by investigating solutions of Hamilton-Jacobi inequalities. In their analysis, the authors follow the lines of the H_∞ optimal control results for linear systems. A certainty equivalence argument is applied to treat a nonlinear version of the H_∞ control problem with imperfect observations. The results of this paper extend a number of recent achievements in this area.

Observer Design for Nonlinear Systems with Discrete-Time Measurements

P. E. Moraal and J. W. Grizzle, *Senior Member, IEEE*

Abstract—This paper focuses on the development of asymptotic observers for nonlinear discrete-time systems. It is argued that instead of trying to imitate the linear observer theory, the problem of constructing a nonlinear observer can be more fruitfully studied in the context of solving simultaneous nonlinear equations. In particular, it is shown that the discrete Newton method, properly interpreted, yields an asymptotic observer for a large class of discrete-time systems, while the continuous Newton method may be employed to obtain a global observer. Furthermore, it is analyzed how the use of Broyden's method in the observer structure affects the observer's performance and its computational complexity. An example illustrates some aspects of the proposed methods; moreover, it serves to show that these methods apply equally well to discrete-time systems and to continuous-time systems with sampled outputs.

I. INTRODUCTION

A. General

THE need to study state estimators (observers) for dynamical systems is, from a control point of view, well understood by now. For the class of finite-dimensional, time-invariant linear systems, a solution to the observer problem has been known since the mid 1960's: the observer incorporates a copy of the system and uses output injection to achieve an exponentially decaying error dynamics. For the class of continuous-time nonlinear systems, the reader is referred to [38], [39] and the references therein for a summary of the theory up to 1986. More recent developments include the work of Krener *et al.* [23] on higher-order approximations for achieving a linearizable error dynamics. Tsinias in [37] has proposed (nonconstructive) existence theorems on nonlinear observers via Lyapunov techniques. Gauthier *et al.* [12] and Deza *et al.* [9] show how to construct high-gain, extended Luenberger- and Kalman-type observers for a class of nonlinear continuous-time systems. Tomambè [36] and Nicosia *et al.* [31] have proposed a continuous-time version of Newton's algorithm as a method for computing the inverse kinematics of robots; moreover, the latter paper also presents a symbiotic relationship in general between asymptotic observers and

nonlinear map inversion. Finally, Michalska and Mayne [27] have used a dual form of moving horizon control to construct observers for nonlinear systems.

Less attention has been focused on the observer problem for discrete-time systems. It was shown in [6] that certain properties, like observer error linearizability [22], are not inherited from the underlying continuous-time system. Moreover, the class of continuous-time systems that admit approximate solutions to the observer error linearization problem for their exact discretizations with sampling time T in an open interval is limited to the class of nonlinear systems that are approximately state-equivalent to a linear system and hence is very restricted [5]. These results motivated the search for a structurally more robust approach to the observer problem. In Section II, it is argued that instead of trying to imitate the linear observer theory, the nonlinear observer problem should be studied in the context of solving sets of simultaneous nonlinear equations. This viewpoint is supported by showing in Section III that the discrete Newton method, properly interpreted, yields an asymptotic observer for a large class of discrete-time systems. In Section IV, a relationship between this observer and the well-known extended Kalman filter is established. As an extension to the result from Section III, it is shown in Section V, that the continuous Newton method may be used to obtain a global exponential observer. Section VI addresses an alternative to using Newton's method in the observer design—namely, Broyden's method—in the case that computational efficiency is an important issue. Finally, an example will illustrate the theory presented herein.

Some of the results reported here have previously appeared in [16], [17], and [28]. Extensions to the case of singularly perturbed discrete-time systems have been presented by Shouse and Taylor [33].

B. Notation and Terminology

Consider a continuous-time system

$$\Sigma_c: \begin{aligned} \dot{x} &= f(x, u) \\ y &= h(x, u) \end{aligned} \quad (1)$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$. Its sampled-data representation, obtained by holding the input constant over half open intervals $[kT, (k+1)T]$ and measuring the output at times kT , will be denoted

$$\Sigma(T): \begin{aligned} x_{k+1} &= F_T(x_k, u_k) \\ y_k &= h(x_k, u_k) \end{aligned} \quad (2)$$

Manuscript received April 23, 1993; revised May 15, 1994. Recommended by Associate Editor, W. P. Dayawansa. This work was supported in part by National Science Foundation Contract NSF ECS-88-96136.

P. E. Moraal was with Department of Electrical Engineering and Computer Science, University of Michigan at Ann Arbor, MI and is now with Ford Motor Company's Research Laboratory, Dearborn, MI, 48121-2053 USA.

J. W. Grizzle is with the Department of Electrical Engineering and Computer Science, University of Michigan at Ann Arbor, MI 48109-2122 USA.

IEEE Log Number 9408270.

where $x_k := x(kT)$, $y_k := y(kT)$, and $u_k := u(kT)$. The symbol " $:=$ " means that the object on the left is defined to be equal to the object on the right; the reverse holds for " $=$."

It is worth noting that if (A, B, C, D) are the matrices describing the Jacobian linearization of (1) around a given equilibrium point, then $(\exp(AT), \int_0^T \exp(A\tau) B d\tau, C, D)$ are the corresponding matrices for (2) about the same equilibrium point [14]. Consequently, if the linearization of (1) is controllable and/or observable, the same will be true of the linearization of (2) for "almost all" T [35].

A discrete-time system will be denoted as

$$\Sigma: \begin{aligned} x_{k+1} &= F(x_k, u_k) \\ y_k &= h(x_k, u_k) \end{aligned} \quad (3)$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^p$. It is convenient to let $F^u(x) := F(x, u)$ and $h^u(x) := h(x, u)$ so that things like $F(F(x, u_1), u_2)$ and $h(F(x, u_1), u_2)$ can be written as $F^{u_2} \circ F^{u_1}(x)$ and $h^{u_2} \circ F^{u_1}(x)$ respectively, where " \circ " denotes composition.

In the sequel, we will often be dealing with a set of N consecutive measurements or controls; these will be denoted as

$$Y_{[k-N+1, k]} := \begin{bmatrix} y_{k-N+1} \\ \vdots \\ y_k \end{bmatrix}, \quad U_{[k-N+1, k]} := \begin{bmatrix} u_{k-N+1} \\ \vdots \\ u_k \end{bmatrix}. \quad (4)$$

If N is fixed and clearly understood, then the abbreviations Y_k and U_k will be employed, so that certain formulas will be easier to read.

To a discrete-time system Σ , we associate an N -lifted system (see [11]), Σ^N , by block processing the measurements and controls over a window of N sampling instances. Specifically, fix N and let $\tilde{Y}_j := Y_{Nj} = Y_{[N(j-1)+1, Nj]}$, $\tilde{U}_j := U_{Nj} = U_{[N(j-1)+1, Nj]}$, and $\tilde{x}_j := x_{Nj}$. Write out \tilde{U}_j in terms of its vector components as $\tilde{U}_j = \text{col}(\tilde{u}_j^1, \dots, \tilde{u}_j^N)$ where $\tilde{u}_j^i := u_{N(j-1)+i}$. Let

$$\Phi(\tilde{x}, \tilde{U}) := F^{\tilde{u}^N} \circ \dots \circ F^{\tilde{u}^1}(\tilde{x}) \quad (5)$$

and

$$H(\tilde{x}, \tilde{U}) = \begin{bmatrix} h^{\tilde{u}^1}(\tilde{x}) \\ \vdots \\ h^{\tilde{u}^N} \circ F^{\tilde{u}^{N-1}} \circ \dots \circ F^{\tilde{u}^1}(\tilde{x}) \end{bmatrix}. \quad (6)$$

The N -lifted system is defined to be

$$\begin{aligned} \tilde{x}_{j+1} &= \Phi(\tilde{x}_j, \tilde{U}_j) \\ \Sigma^N: \quad \tilde{Y}_j &= H(\tilde{x}_j, \tilde{U}_j) \end{aligned} \quad (7)$$

Note that its dynamics is nothing more than the dynamics of (3) iterated N -times. The state of (7) is the state of (3) at the beginning of each "window" of length N , and Φ simply describes how the state evolves from window to window. The representation (7) can be termed "multirate" because, if (3)

arises from a continuous-time system (1), (7) can be obtained directly from (1) by sampling the inputs and outputs N -times faster than the state; in other words, $\tilde{x}_j := x(jNT)$, $\tilde{u}_j^i = u((j-1)NT + iT)$, etc. More generally, one could sample the inputs and outputs at different rates, or even some input components at faster rates than others, but we will not pursue this here. Throughout this paper, the notation $\|\cdot\|$ will be used to denote both a vector norm and the corresponding induced operator norm. Finally, we recall that if $g: \mathbb{R}^{r_1} \rightarrow \mathbb{R}^{r_2}$ is at least once continuously differentiable, its rank at a point $x_0 \in \mathbb{R}^{r_1}$ is the rank of its Jacobian matrix at x_0 , [4] that is, $\text{rank} [\frac{\partial g}{\partial x}(x_0)]$.

II. OBSERVERS FOR SMOOTH DISCRETE-TIME NONLINEAR SYSTEMS

A. General

Consider a discrete-time system on \mathbb{R}^n

$$\Sigma: \begin{aligned} x_{k+1} &= F(x_k, u_k) \\ y_k &= h(x_k, u_k) \end{aligned} \quad (8)$$

A second system

$$\begin{aligned} z_{k+1} &= \Gamma(z_k, y_k, u_k) \\ \hat{x}_k &= \eta(z_k, y_k, u_k) \end{aligned} \quad (9)$$

with $z_k \in \mathbb{R}^l$, some $l \geq 0$, is an asymptotic observer [25] for (8) if it satisfies: A) $\forall x_1 \in \mathbb{R}^n, \forall u_k \in \mathbb{R}^m, \exists z_1 \in \mathbb{R}^l$ such that $\hat{x}_k = x_k$ for all $k \geq 2$, and B) $\forall x_1 \in \mathbb{R}^n, \forall u_k \in \mathbb{R}^m, z_1 \in \mathbb{R}^l, \lim_{k \rightarrow \infty} \|\hat{x}_k - x_k\| = 0$. If the read-out map η in (9) is the identity, $\hat{x}_k = z_k$, then (9) is called an identity observer [25]; if the convergence of \hat{x} to x is exponential, then (9) is called an exponential observer.

For later use, the observer (9) will be said to be dead-beat of order d , if, upon writing $\Gamma(z_k, y_k, u_k) =: \Gamma^{y_k, u_k}(z_k)$ and $\eta(z_k, y_k, u_k) =: \eta^{y_k, u_k}(z_k)$, then

$$\eta^{y_d, u_d} \circ \Gamma^{y_{d-1}, u_{d-1}} \circ \dots \circ \Gamma^{y_1, u_1}(z_1) = x_d \quad (10)$$

independently of the particular observer initial condition z_1 , where x_d is the state of (8) at time d . It is remarked that dead-beat observers are of interest for stabilization problems, because, if $u_k = \alpha(x_k)$ is a stabilizing feedback for (8), then $u_k = \alpha(\hat{x}_k)$ will always result in an internally stable closed-loop system whenever the observer (9) has the dead-beat property. This is one of the rare instances of a nonlinear separation principle.

All the above has been stated in a global fashion. Let us note that there are at least two ways of localizing the concept of an observer. The first is essentially infinitesimal: one guarantees the existence of open neighborhoods \mathcal{O}_x and \mathcal{O}_z of the origin of (8) and (9), respectively, and an open neighborhood of controls \mathcal{O}_u such that A) and B) hold as long as $z \in \mathcal{O}_z$ and $\forall k \geq 1, u_k \in \mathcal{O}_u$ and $x_k \in \mathcal{O}_x$. The work on observers with linearizable error dynamics [6], [20]–[24], for instance, falls into this category. A second way to localize the concept

ould be called (S, \mathcal{V}) -quasilocal: one is given subsets S and \mathcal{V} , of the state space of (8) and of its controls, respectively, having the property that, for every initial point $x_1 \in S$, there exists an open subset $\mathcal{O}_z(x_1)$ of the state space of (9), such that A) and B) hold as long as $z_1 \in \mathcal{O}_z(x_1)$ and $\forall k \geq 1$, $r_k \in S$, and $u_k \in \mathcal{V}$. In other words, for the case of identity observers, instead of guaranteeing the existence of an open set about the origin of the product state space $\mathbb{R}^n \times \mathbb{R}^n$ where everything works, one is assuring the existence of an open set about the diagonal of $\mathbb{R}^n \times \mathbb{R}^n$, whose projection onto the r -coordinate contains S .

In the following, an approach to the construction of observers for discrete-time systems is developed. The authors' perspective was influenced by the work of Aeyels [1], [2], Fitts [10], Glad [13], and the multi-rate time sampling results of [14].

B Dead-Beat Observers

Consider once again the system Σ (8) and let $Y_{[1,N]}$ denote a vector of N consecutive measurements

$$Y_{[1,N]} = \begin{pmatrix} h^{u_1}(x) \\ h^{u_2} \circ F^{u_1}(x) \\ \vdots \\ h^{u_N} \circ F^{u_{N-1}} \circ \dots \circ F^{u_1}(x) \end{pmatrix} =: H(x, U_{[1,N]}) \quad (11)$$

Σ is said to be N -observable¹ [1], [32], [35] at a point $x \in \mathbb{R}^n$, $N \geq 1$, if there exists an N -tuple of controls $U_{[1,N]} = \text{col}(u_1, \dots, u_N) \in (\mathbb{R}^m)^N$ such that x is the unique solution of the set of equations

$$\bar{Y}_{[1,N]} = H(x, U_{[1,N]}) \quad (12)$$

where

$$\bar{Y}_{[1,N]} = H(\bar{x}, U_{[1,N]}). \quad (13)$$

The system is uniformly N -observable if the mapping

$$H^*: \mathbb{R}^n \times (\mathbb{R}^m)^N \rightarrow (\mathbb{R}^p)^N \times (\mathbb{R}^m)^N \quad (14)$$

by $(x, U_{[1,N]}) \mapsto (H(x, U_{[1,N]}), U_{[1,N]})$ is injective; it is locally uniformly N -observable with respect to $\mathcal{O} \subset \mathbb{R}^n$ and $\mathcal{U} \subset (\mathbb{R}^m)^N$ if H^* restricted to $\mathcal{O} \times \mathcal{U}$ is injective.

Whenever Σ is uniformly N -observable, the system of equations

$$Y_{[k-N+1,k]} = H(x_{k-N+1}, U_{[k-N+1,k]}) \quad (15)$$

can be, for each N applied inputs $U_{[k-N+1,k]}$, uniquely solved for x_{k-N+1} , and the current state x_k obtained by

$$x_k = \Phi^{U_{[k-N,k-1]}}(x_{k-N+1}). \quad (16)$$

¹ N refers to the minimum number of measurements needed to recover the state. In [2], Aeyels shows that, "generically," N can be taken to be $2n + 1$.

This constitutes an order N dead-beat observer for Σ , [14]. Conversely, suppose that (9) is a dead-beat observer of order N . Then

$$\eta^{y_N, u_N} \circ \Gamma^{y_{N-1}, u_{N-1}} \circ \dots \circ \Gamma^{y_1, u_1}(z) = r_N \quad (17)$$

for all $z \in \mathbb{R}^l$; thus the left-hand side of (17) does not depend on z and is a solution to (15)–(16). This shows that constructing a dead-beat observer of order N is equivalent to left-inverting (15) and composing the result with the right-hand side of (16). In a similar vein, an asymptotic (nondead-beat) observer can be thought of as constructing a solution to (15)–(16) as $N \rightarrow \infty$. Clearly, for nonlinear systems, insisting that this can be done in closed-form is very restrictive. It is therefore natural to formulate an extended concept of an observer as a possibly implicitly defined dynamical system, involving successive approximation routines, logical variables and/or lookup tables to dynamically "estimate" the state of a deterministic nonlinear system [16]. This perspective will be further pursued in the next section where Newton's algorithm is interpreted as a nonlinear observer (9).

Before doing so, however, let us first tie in the notion of a dead-beat observer with the observer error linearization approach [6], [20]–[24]. For simplicity of exposition, suppose that (8) does not have any inputs. One seeks a (locally defined) coordinate transformation $x = T(\bar{x})$ in which (8) takes the form

$$\begin{aligned} x_{k+1} &= Ax_k + \psi(y_k) \\ y_k &= C\bar{x}_k \end{aligned} \quad (18)$$

where the pair (A, C) is observable. This gives a family of infinitesimally-local observers

$$\begin{aligned} \hat{z}_{k+1} &= (A - KC)\hat{z}_k + \psi(y_k) + Ky_k \\ \hat{x}_k &= T^{-1}(\hat{z}_k). \end{aligned} \quad (19)$$

Letting $e_k := x_k - \hat{x}_k$ yields

$$e_{k+1} = (A - KC)e_k. \quad (20)$$

Choosing K to place the eigenvalues of $(A - KC)$ at zero makes $\hat{\Sigma}$ into a dead-beat observer of order n . In other words, the ability to achieve a linear error dynamics (20) implies the explicit knowledge of a left-inverse to (12).

III. NEWTON'S ALGORITHM AS AN OBSERVER

Consider again the system Σ , (8). It is said to satisfy the N -observability rank condition with respect to $\mathcal{O} \subset \mathbb{R}^n$ and $\mathcal{U} \subset (\mathbb{R}^m)^N$ if $H^*: \mathcal{O} \times \mathcal{U} \rightarrow (\mathbb{R}^p)^N \times (\mathbb{R}^m)^N$ is an immersion [35]; that is, it has rank $n + Nm$ at each point of $\mathcal{O} \times \mathcal{U}$ (recall that H^* was defined in (14)). Note that Σ is N -observable and satisfies the N -observability rank condition with respect to \mathcal{O} and \mathcal{U} if, and only if, $H^*: \mathcal{O} \times \mathcal{U} \rightarrow (\mathbb{R}^p)^N \times (\mathbb{R}^m)^N$ is an injective immersion;² this is in turn equivalent to: for each $U_{[1,N]} \in \mathcal{U}$, $H(\cdot, U_{[1,N]}): \mathcal{O} \rightarrow (\mathbb{R}^p)^N$ is an injective immersion.

²That is, an embedding [4].

Newton's algorithm for

$$Y_{[k-N+1,k]} - H(x_{k-N+1}, U_{[k-N+1,k]}) = 0 \quad (21)$$

is

$$\xi^{i+1} = \xi^i + \left[\frac{\partial H}{\partial x}(\xi^i, U_{[k-N+1,k]}) \right]^{-1} \cdot (Y_{[k-N+1,k]} - H(\xi^i, U_{[k-N+1,k]})) \quad (22)$$

where, for simplicity, it has been assumed that the set of (21) is square; in the case that there are more equations than states, the inverse in (22) should be replaced by a pseudo-inverse [26, p. 309], [8, pp. 222–224]. The standard convergence theorem for this algorithm can be found in [26]. For the moment, assume that $U_{[k-N+1,k]}$ is fixed, and let $H(x) = H(x, U_{[k-N+1,k]})$ for this fixed value of U .

Theorem 3.1 [26]: Suppose that H is twice differentiable and that $\|\frac{\partial^2 H}{\partial x^2}(x)\| \leq K$ for $x \in \mathbb{R}^n$; suppose there is a point $\xi^0 \in \mathbb{R}^n$ such that $P_0 := \frac{\partial H}{\partial x}(\xi^0)$ is invertible with $\|P_0^{-1}\| \leq \beta_0$ and $\|P_0^{-1}(Y_k - H(\xi_0))\| \leq \eta_0$. Under these conditions, if the constant $h_0 = \beta_0 \eta_0 K < 1/2$, then the sequence ξ^i generated by (22) exists for all $i \geq 0$ and converges to a solution of (21). If instead $\|\frac{\partial^2 H}{\partial x^2}(x)\| \leq K$ only in a neighborhood B of ξ_0 with radius

$$r \geq \frac{1}{h_0} (1 - \sqrt{1 - 2h_0}) \eta_0 \quad (23)$$

then the successive approximations generated by Newton's algorithm remain within this neighborhood and converge to a solution of (21).

The most interesting point is that Theorem 3.1 gives an estimate of how good the initial estimate of x_k should be before a few iterations of (22) will generate better estimates. In this regard, the quantity $\|\frac{\partial^2 H}{\partial x^2}(x)\|$, which measures the degree of nonlinearity of (21), is seen to be of central importance. For a linear system, $\|\frac{\partial^2 H}{\partial x^2}(x)\| \equiv 0$, and the initial estimate can be arbitrarily poor; when $\|\frac{\partial^2 H}{\partial x^2}(x)\|$ is large, the initial estimate should, in general, be better.

Newton's algorithm is now interpreted as a quasilocal exponential observer. Suppose that N has been fixed; for notational ease, let $Y_k = Y_{[k-N+1,k]}$ be the vector of the last N measurements and similarly let $U_k = U_{[k-N+1,k]}$ be $\text{col}(u_{k-N+1}, \dots, u_k)$ be the vector of the last N controls. Define

$$\Theta^{Y_k, U_k}(\zeta) = \zeta + \left[\frac{\partial H}{\partial x}(\zeta, U_k) \right]^{-1} (Y_k - H(\zeta, U_k)) \quad (24)$$

and let $(\Theta^{Y_k, U_k})^{(d)}(\xi)$ represent $\Theta^{Y_k, U_k}(\xi)$ composed with itself d -times.

Let \mathcal{O} be a subset of \mathbb{R}^n , \mathcal{V} a subset of \mathbb{R}^m , $N \geq 1$ a given integer and $\epsilon > 0$ a positive constant. Denote the complement of \mathcal{O} by $\sim \mathcal{O}$ and define $\text{dist}(x, \sim \mathcal{O}) = \inf\{\|x - y\| : y \in \sim \mathcal{O}\}$, and $\mathcal{O}_{\epsilon/2} = \{x \in \mathcal{O} : \text{dist}(x, \sim \mathcal{O}) \geq \epsilon/2\}$. Finally, define

constants α , β , γ , L , and C by

$$\begin{aligned} \alpha &= \sup \left\{ \left\| \left[\frac{\partial H}{\partial x}(x, U) \right]^{-1} \right\| : x \in \mathcal{O}_{\epsilon/2}, U \in \mathcal{V}^N \right\} \\ \beta &= \sup \left\{ \left\| \frac{\partial H}{\partial x}(x, U) \right\| : x \in \mathcal{O}_{\epsilon/2}, U \in \mathcal{V}^N \right\} \\ \gamma &= \sup \left\{ \left\| \frac{\partial^2 H}{\partial x^2}(x, U) \right\| : x \in \mathcal{O}_{\epsilon/2}, U \in \mathcal{V}^N \right\} \\ L &= \sup \left\{ \left\| \frac{\partial F}{\partial x}(x, u) \right\| : x \in \mathcal{O}_{\epsilon/2}, u \in \mathcal{V} \right\} \\ C &= \frac{1}{2} \sup \left\{ \left\| \frac{\partial^2 \Theta^{Y, U}}{\partial x^2}(x) \right\| : Y = H(x, U), \right. \\ &\quad \left. x \in \mathcal{O}_{\epsilon/2}, U \in \mathcal{V}^N \right\}. \end{aligned}$$

Theorem 3.2: Suppose that the following conditions hold:

- 1) F and h in (8) are at least three times differentiable with respect to x ;
- 2) there exist a bounded subset $\mathcal{O} \subset \mathbb{R}^n$ and a compact subset $\mathcal{V} \subset \mathbb{R}^m$ such that for each $x \in \mathcal{O}$ there exists $u \in \mathcal{V}$ such that $F(x, u) \in \mathcal{O}$ (i.e., \mathcal{O} is controlled-invariant with respect to \mathcal{V}); moreover, the controls are always applied so that $F(x, u) \in \mathcal{O}$;
- 3) there exists an integer $1 \leq N \leq n$ such that the set of equations (11) is
 - a) square,
 - b) uniformly N -observable with respect to \mathcal{O} and \mathcal{V}^N ,
 - c) satisfies the N -observability rank condition with respect to \mathcal{O} and \mathcal{V}^N .

Then, for every $\epsilon > 0$, the constants α , β , γ , L , and C are finite; moreover, whenever

$$\delta \leq \min \left\{ \frac{\epsilon}{4L}, \frac{1}{4\gamma\beta(\alpha)^2 L}, \frac{\epsilon}{8\beta\alpha L}, \frac{1}{2CL} \right\} \quad (25)$$

and

$$d \geq \max\{1, \log_2 \log_2 4L\}, \quad d \in \mathbb{N} \quad (26)$$

then

$$z_{k+1} = (\Theta^{Y_k, U_k})^{(d)}(F(z_k, u_{k-N})) \quad (27)$$

$$\hat{x}_k = F^{u_{k-1}} \circ F^{u_{k-2}} \circ \dots \circ F^{u_{k-N}}(z_k) \quad (28)$$

is a quasilocal, exponential observer for (8) in the sense that. A) if $x_1 \in \mathcal{O}$ and $z_{N+1} = x_1$, then $\hat{x}_k = x_k$ for all $k \geq N+1$ and B), if $x_1 \in \mathcal{O}$, $\|z_{N+1} - x_1\| < \delta$ and for all $k \geq 0$, $\text{dist}(x_k, \sim \mathcal{O}) \geq \epsilon$, then $\|\hat{x}_{k+1} - x_{k+1}\| \leq \frac{1}{2} \|\hat{x}_k - x_k\|$.

The proof may be found in [17]; the basic idea is to view the observer problem as one of solving a sequence of nonlinear inversion problems, each described by (12). Since the set \mathcal{O} is relatively compact and controlled invariant with a compact set of controls, Newton's algorithm can be shown to have a uniform rate of convergence over the entire sequence of problems. The idea then is to iterate long enough on each problem (the parameter d) so that F applied to the solution of the k th problem is a very good initial guess for the $(k+1)$ st problem.

The set \mathcal{O} is assumed to be bounded, but not necessarily small; if it is not controlled-invariant, then only finite time estimates are possible; the same is true of the observer error linearization approach of [20]–[24] (for discrete-time systems, see [6]).

The observer (27)–(28) is coordinate dependent. It is interesting to note that the coordinate transformation approaches, in general, would only favor convergence of (21) if they reduce $\| \frac{\partial^2 H}{\partial x^2}(x, U) \|$. In particular, eliminating low-order polynomial terms in favor of high-order terms will not always accomplish this task.

Remark 3.3:

- In Theorem 3.2, one may take $d = 1$ if, in (25), $\frac{1}{2CL}$ is replaced by $\frac{1}{2CL^2}$.
- Once again, assumption 3-a), that (11) is square, is NOT essential. One could try eliminating certain rows of (11) while still preserving the rank condition 3-c), but this would, more-than-likely, invalidate 3-b). The better alternative is to replace the inverse in (24) with a pseudo-inverse, as in [26, p. 309] or [8, pp. 222–224].
- The observer (27)–(28) bears some resemblance to the iterated extended Kalman filter of [7]. This will formally be established in the next section.
- By modifying the step-size in Newton's algorithm, "globally convergent" versions of the algorithm can be shown to exist. Chapter 6 of [8] presents this very nicely from a numerical analytic viewpoint. A different way of "globalizing" the algorithm is to systematically produce a good point at which to initialize it. This is discussed in [15], [16]. In Section V, the continuous Newton method will be shown to yield a global version of the above observer.
- It is often pointed out, and in [29] shown to be a valid practical concern, that the evaluation of the Jacobian $\frac{\partial H}{\partial x}$, be it explicitly or using finite difference approximations, may be computationally very expensive or even prohibitive. Modified Newton methods have been proposed, in which the Jacobian is not explicitly evaluated at every step, but updated iteratively without requiring additional function evaluations. Section VI explores the consequences of using Broyden's method instead of Newton's in the observer (27)–(28).

IV. RELATION BETWEEN KALMAN FILTERS AND NEWTON OBSERVERS

To show how the Newton observer is related to the extended Kalman filter, we will consider an invertible, autonomous discrete-time system

$$\begin{aligned} x_{k+1} &= F(x_k), & x_k &\in \mathbb{R}^n \\ y_k &= h(x_k), & y_k &\in \mathbb{R}^p \end{aligned} \quad (29)$$

in which we replace the output map h by the extended output map H , defined in the following manner

$$Y_k = \begin{pmatrix} y_{k-N+1} \\ \vdots \\ y_{k-1} \\ y_k \end{pmatrix} = \begin{pmatrix} h \circ F^{-(N-1)}(x_k) \\ \vdots \\ h \circ F^{-1}(x_k) \\ h(x_k) \end{pmatrix} =: H(x_k). \quad (30)$$

Assume that the above system satisfies the N -observability rank condition and is N -observable; furthermore, assume that H is a square map, i.e., $H: \mathbb{R}^n \rightarrow \mathbb{R}^n$. A common way

to construct an observer for system (29)–(30) is to apply the extended Kalman filter to the associated noisy system, i.e., the system with added artificial noise processes

$$\begin{aligned} z_{k+1} &= F(z_k) + Nw_k \\ \xi_k &= H(z_k) + Rv_k \end{aligned} \quad (31)$$

where v_k and w_k are assumed to be jointly Gaussian and mutually independent random processes with zero mean and unit variance. The extended Kalman filter for this system is given by the following equations:

measurement update

$$\begin{aligned} \hat{x}_k &= \hat{x}_k^- + K_k(\xi_k - H(\hat{x}_k^-)), \\ Q_k^{-1} &= (Q_k^-)^{-1} + H_k^T(RR^T)^{-1}H_k \end{aligned}$$

time update

$$\begin{aligned} \hat{x}_{k+1}^- &= F(\hat{x}_k), \\ Q_{k+1}^- &= A_k Q_k A_k^T + NN^T \end{aligned}$$

where

$$\begin{aligned} K_k &= Q_k^- H_k^T (H_k Q_k^- H_k^T + RR^T)^{-1}, \\ A_k &= \frac{\partial F}{\partial x}(\hat{x}_k), \\ H_k &= \frac{\partial H}{\partial x}(\hat{x}_k^-) \end{aligned}$$

and N , R , and Q_0 are the design parameters. Let us choose $N = \mu I$ and $R = \varepsilon I$, and consider the equations for the error covariance Q_k^- and the observer gain K_k

$$\begin{aligned} Q_{k+1}^- &= A_k \left((Q_k^-)^{-1} + \frac{1}{\varepsilon^2} H_k^T H_k \right)^{-1} A_k^T + \mu^2 I \\ &= \varepsilon^2 A_k (\varepsilon^2 (Q_k^-)^{-1} + H_k^T H_k)^{-1} A_k^T + \mu^2 I \\ K_k &= Q_k^- H_k^T (H_k Q_k^- H_k^T + \varepsilon^2 I)^{-1}. \end{aligned}$$

Given any positive definite Q_0^- , the update equation for Q_k^- in the limit as $\varepsilon \rightarrow 0$ is given by

$$Q_{k+1}^- = \mu^2 I.$$

Substituting this in the equation for K_k and letting $\varepsilon \rightarrow 0$ gives

$$\begin{aligned} K_k &= \mu^2 H_k^T (H_k \mu^2 H_k^T)^{-1} \\ &= H_k^T (H_k H_k^T)^{-1} \\ &= H_k^{-1} \end{aligned}$$

which is valid since, given the observability conditions, H_k is invertible. The extended Kalman filter equations are then given by

$$\hat{x}_k = \hat{x}_k^- + H_k^{-1}(Y_k - H(\hat{x}_k^-)) \quad (32)$$

$$\hat{x}_{k+1}^- = F(\hat{x}_k) \quad (33)$$

which is exactly the Newton observer for system (29) with one Newton iteration per time step.

In [3], it was recently shown that the measurement update equations for \hat{x}_k in the iterated extended Kalman filter are exactly those arising from the minimization problem

$$\min_{x_k} \left((Y_k - H(\hat{x}_k^-))^T R^{-1} (Y_k - H(\hat{x}_k^-)) + (x_k - \hat{x}_k^-)^T (Q_k^-)^{-1} (x_k - \hat{x}_k^-) \right) \quad (34)$$

when a Gauss-Newton method (an approximate Newton method) is used with \hat{x}_k^- as initial guess. This shows that the covariance matrices R and Q_k^- may be interpreted as weights on the norms in the output space and state space, respectively. It remains presently unclear, however, how the update equations for the covariance matrix Q_k^- in the Kalman filter can be given a meaningful interpretation in terms of updating the weighting matrix in the above minimization problem after every iteration.

It must be pointed out that the extended Kalman filter, although commonly used as a nonlinear observer, had not been actually proven to be a convergent asymptotic nonlinear observer until recently. In [34], it is shown that, under suitable observability conditions, if the state evolves in a compact set and the Kalman filter is initialized close enough to the true state, then the error covariance matrices Q_k^- and $(Q_k^-)^{-1}$ remain bounded, and the observer error goes to zero exponentially. A proof of convergence for the continuous Kalman filter with a special choice for the initial error covariance matrix is given in [9].

V. CONTINUOUS NEWTON METHOD AS A GLOBAL OBSERVER

In the remaining sections, we will, for notational simplicity and without loss of generality, restrict ourselves to invertible and autonomous systems. The results that are obtained can without any difficulty be extended to noninvertible systems and/or systems with inputs (see example section). Consider once again the discrete-time system

$$\begin{aligned} \hat{x}_{k+1} &= F(x_k), \quad x_k \in \mathbb{R}^n \\ y_k &= h(x_k), \quad y_k \in \mathbb{R}^p \end{aligned} \quad (35)$$

with the extended output map H , defined in the following manner

$$Y_k = \begin{pmatrix} y_{k-N+1} \\ \vdots \\ y_{k-1} \\ y_k \end{pmatrix} = \begin{pmatrix} h \circ F^{-(N-1)}(x_k) \\ \vdots \\ h \circ F^{-1}(x_k) \\ h(x_k) \end{pmatrix} =: H(x_k). \quad (36)$$

Assume that H is a square map.³ The discrete Newton method with step size h_1 for solving $Y_k - H(x) = 0$ is

$$z_k^{i+1} = z_k^i + h_1 J(z_k^i)^{-1} (Y_k - H(z_k^i)) \quad (37)$$

where $J(z) := \frac{\partial H}{\partial x}(z)$. If we consider this equation with an infinitesimally small step size, we obtain the following differential equation

$$\dot{z}_k = J(z_k)^{-1} (Y_k - H(z_k)), \quad z_k(0) = z_k^0 \quad (38)$$

which is referred to as the continuous Newton method; the right hand side of (38) is commonly referred to as (gradient) Newton flow [19]. The stability of Newton flows has been studied extensively (see [19], [40] and references therein).

We now construct a global asymptotic high-gain hybrid observer for the system (35), by interpreting (38) as an observer

for that system. Assume that the time interval between x_k and x_{k+1} is T , i.e., $x_k = x(kT)$. We thus obtain the following hybrid system-observer

$$x_{k+1} = F(x_k) \quad (39)$$

$$Y_k = H(x_k)$$

$$\begin{aligned} \dot{z}_k &= K J(z_k)^{-1} [Y_k - H(z_k)]; \quad z_k(kT) = F(z_{k-1}(kT)) \\ \hat{x}_k &= z_k((k+1)T) \end{aligned} \quad (40)$$

where K is a positive scalar, to be determined later.

Theorem 5.1. Assume that (35) is uniformly N -observable and satisfies the N -observability rank condition with respect to \mathbb{R}^n . Suppose that

$$\begin{aligned} L &:= \sup_{r \in \mathbb{R}^n} \left\| \frac{\partial F}{\partial r}(r) \right\| < \infty; \\ \beta &:= \sup_{r \in \mathbb{R}^n} \left\| \frac{\partial F}{\partial r}(r) \right\| < \infty; \\ \alpha &:= \sup_{x \in \mathbb{R}^n} \left\| \left(\frac{\partial H}{\partial x}(x) \right)^{-1} \right\| < \infty \end{aligned} \quad (41)$$

If $K \geq \frac{1}{T} \log(2L\alpha/\beta)$, then (39) is a global asymptotic observer for (35).

The proof for this and the next result can be found in [28]. In general, the quantities $\|\frac{\partial F}{\partial r}\|$, $\|\frac{\partial H}{\partial x}\|$, and $\|\frac{\partial F}{\partial x}\|$ will not be uniformly bounded on \mathbb{R}^n , nor will the system (35) be globally observable. For these cases, we can still obtain a nonlocal convergence result for the observer (40).

Proposition 5.2. Suppose the following conditions hold:

- 1) F and h are at least once continuously differentiable;
- 2) \exists compact $\mathcal{O} \subset \mathbb{R}^n$ such that:
 - a) (35) is N -observable with respect to \mathcal{O} ;
 - b) (35) satisfies the N -observability rank condition with respect to \mathcal{O} ;
- 3) $\exists \mu > 0$ such that $\mathcal{O}_\mu := \{x \in \mathcal{O} \mid \inf_{y \in \mathbb{R}^n \setminus \mathcal{O}} \|x - y\| > \mu\}$ is F -invariant and nonempty.

Then, if $K \geq \frac{1}{T} \log(2L\alpha/\beta)$, if $x_0 \in \mathcal{O}_\mu$ and if $\|z_0(0) - x_0\| < \frac{\mu}{\beta\alpha}$, it follows that $\lim_{k \rightarrow \infty} \|\hat{x}_k - x_k\| = 0$.

Remark 5.3 If no *a priori* information is known about the initial state of the system, one may, for lack of a better alternative, initialize the observer at the origin. Suppose \mathcal{O} contains the closed ball centered at the origin, with radius R $B(0, R)$. Then the previous analysis showed that convergence can be guaranteed if $\|\hat{x}_0 - x_0\| = \|x_0\| \leq \frac{R}{1+\beta\alpha}$ or, what may provide a more practical estimate, if $\|H(0) - Y_{N-1}\| \leq \frac{\beta R}{1+\beta\alpha}$. Most global modifications of the discrete Newton method are based on choosing a proper stepsize (e.g., Armijo stepsize procedures) and/or search direction (e.g., trust region updates) [8], [30], such as to assure a decrease in the term $\|Y_k - H(z^i)\|$ in the i th step of the algorithm. Obviously, in terms of a connected region of convergence, one cannot do better than allowing an infinite number of iterations, each taking infinitesimally small steps in a guaranteed descent direction, i.e., the continuous Newton method.

³This seems to be a crucial assumption

VI. BROYDEN'S METHOD

In the previous section, it was seen that a large region of convergence of the Newton-observer could be guaranteed if one used the hybrid form presented in (39). To implement the hybrid Newton-observer, however, a closed-form expression for the inverse of the Jacobian matrix would be necessary; this does not seem very realistic. Moreover—as mentioned in Section III—even in the discrete Newton algorithm, the computational complexity of repeated Jacobian evaluations might prove prohibitive in practice. This provides sufficient motivation to pursue methods for approximating the Jacobian and the investigation of the associated convergence properties. Here, we will look at Broyden's method, as it is among the most popular of such approximation schemes [8], [30].

Broyden's method for solving a system of nonlinear equations $P(x) = 0$ is a modified Newton method in which the Jacobian is not calculated exactly at each step, but rather iteratively approximated using secant updates. In contrast to, for example, finite difference approximations, no additional function evaluations are required. Define $J(x) = \frac{\partial P}{\partial x}(x)$. Broyden's method for solving $P(x) = 0$ is [8]: Given x^0 , the initial guess for \bar{x} , where $P(\bar{x}) = 0$, and A^0 , the initial approximation for the Jacobian of P at x^0

$$\text{solve } A' s' = -P(x') \text{ for } s' \quad (42a)$$

$$x'^{+1} := x' + s' \quad (42b)$$

$$v' := P(x'^{+1}) - P(x') \quad (42c)$$

$$A'^{+1} := A' + \frac{(v' - A' s')(s')^T}{(s')^T s'}. \quad (42d)$$

The update A' for the Jacobian has the property of bounded deterioration: even though it may not converge to $J(x)$, it deteriorates slowly enough for one to still be able to prove convergence of $\{x^i\}$ to \bar{x} .

Conditions for convergence are the same as those of Newton's method, with the additional requirement that, with $P(\bar{x}) = 0$, not only the initial guess x^0 be sufficiently close to \bar{x} , but that also A^0 be sufficiently close to $J(x)$. For Broyden's method, local superlinear convergence can be proven. Newton's method, on the other hand, exhibits local quadratic convergence.

In the following we will show that, in general, Broyden's method alone cannot successfully be used in the Newton-observer; occasional recalculation of the exact Jacobian seems to remain necessary, basically because the property of bounded deterioration no longer holds when Broyden's method is applied to a sequence of problems: $P_k(x) = 0$. For the class of slow-varying or weakly nonlinear systems, however, this approach can substantially reduce the computational complexity.

In the observer problem, at each step k we want to solve

$$P_k(x_k) := H(x_k) - Y_k = 0. \quad (43)$$

Note that this sequence of equations has a special structure: $\frac{\partial P_{k+1}}{\partial x} \equiv \frac{\partial P_k}{\partial x}$ and their solutions are related by $x_{k+1} = F(x_k)$. Assume for simplicity of exposition that $H: \mathbb{R}^n \rightarrow \mathbb{R}^n$. Let

\mathcal{O} be a subset of \mathbb{R}^n such that the following quantities are finite

$$L := \sup_{x \in \mathcal{O}} \left\| \frac{\partial F}{\partial x}(x) \right\| < \infty$$

$$\beta := \sup_{x \in \mathcal{O}} \left\| \frac{\partial H}{\partial x}(x) \right\| < \infty$$

$$\alpha := \sup_{x \in \mathcal{O}} \left\| \left(\frac{\partial H}{\partial x}(x) \right)^{-1} \right\| < \infty$$

$$\gamma := \sup_{x \in \mathcal{O}} \left\| \frac{\partial^2 H}{\partial x^2}(x) \right\| < \infty.$$

Given the d th iterate in the k th problem, x_k^d and A_k^d , being approximations for x_k and $J(x_k)$, respectively, the initial guess for the $(k+1)$ st problem is taken as $x_{k+1}^0 = F(x_k^d)$. In terms of (42a)–(42d), this defines \bar{s}_k and \bar{v}_k , and hence A_{k+1}^0 , the initial approximation for the Jacobian at x_{k+1}^0 as

$$A_{k+1}^0 = A_k^d + \frac{(\bar{v}_k - A_k^d \bar{s}_k) \bar{s}_k^T}{\bar{s}_k^T \bar{s}_k} \quad (44a)$$

where

$$\bar{s}_k = F(x_k^d) - x_k^d \quad (44b)$$

$$\bar{v}_k = P_{k+1}(x_{k+1}^0) - P_{k+1}(x_k^d) \quad (44c)$$

$$x_{k+1}^0 = F(x_k^d). \quad (44d)$$

Note that this update is the same as in Broyden's method, (42d), except that s_k is determined by (44b), which is a consequence of switching from the k th to the $(k+1)$ -st problem in the sequence $\{P_k(x) = 0\}$.

To develop bounds on the approximation error, we will need the following lemma from [8].

Lemma 6.1 (Bounded Deterioration): Let $D \subseteq \mathbb{R}^n$ be open and convex; $x^i, x^{i+1} \in D, x^i \neq \bar{x}$. Let $A^i \in \mathbb{R}^{n \times n}$ and let A^{i+1} be defined by (42a)–(42d). Assume that J is such that $\exists \gamma < \infty$ verifying

$$\|J(x) - J(\bar{x})\| \leq \gamma \|x - \bar{x}\| \quad \forall x \in D. \quad (45)$$

Then, for either the Frobenius or the l_2 -matrix norm

$$\|A^{i+1} - J(\bar{x})\| \leq \|A^i - J(\bar{x})\| + \frac{1}{2} \gamma (\|x^{i+1} - \bar{x}\|_2 + \|x^i - \bar{x}\|_2). \quad (46)$$

To begin with, let x_k and x_{k+1} be such that $P_k(x_k) = H(x_k) - Y_k = 0$ and $P_{k+1}(x_{k+1}) = H(x_{k+1}) - Y_{k+1} = 0$. Let A_k^d and x_k^d be given and determine A_{k+1}^0 and x_{k+1}^0 by (44a)–(44d). Furthermore, define $E_k^i := A_k^i - J(x_k)$ and $e_k^i := x_k^i - x_k$. A bound on the error $\|E_{k+1}^0\|$ can then be derived to be

$$\begin{aligned} \|E_{k+1}^0\| &\leq \|E_k^d\| + \gamma \|x_k - x_{k+1}\| \\ &\quad + \frac{1}{2} \gamma (\|e_k^d\| + \|e_{k+1}^0\| + \|x_{k+1} - x_k\|) \\ &\leq \|E_k^d\| + \frac{1}{2} \gamma (1 + L) \|e_k^d\| + \frac{3}{2} \gamma \|x_k - x_{k+1}\|. \end{aligned} \quad (47)$$

From Lemma 6.1 we get the following

$$\|E_k^d\| \leq \|E_k^0\| + \frac{1}{2} \gamma \left(\|e_k^0\| + 2 \sum_{i=1}^d \|e_k^i\| \right). \quad (48)$$

From the superlinear convergence of Broyden's method, it follows that, with $\|E_k^0\|$ sufficiently small, for any $0 < c < 1$, the following holds: provided that $\|e_k^0\|$ is sufficiently small

$$\|e_k^{i+1}\| \leq c\|e_k^i\| \quad \forall i \geq 0. \quad (49)$$

Now we can write (48) as

$$\begin{aligned} \|E_k^d\| &\leq \|E_k^0\| + \frac{1}{2}\gamma \left(\|e_k^0\| + 2 \left(\frac{1-c^{d+1}}{1-c} - 1 \right) \|e_k^0\| \right) \\ &\leq \|E_k^0\| + \frac{1}{2}\gamma \|e_k^0\| \left(\frac{1+c-2c^{d+1}}{1-c} \right) \end{aligned} \quad (50)$$

and (47) becomes

$$\begin{aligned} \|E_{k+1}^0\| &\leq \|E_k^0\| + \frac{3}{2}\gamma \|x_k - x_{k+1}\| \\ &\quad + \frac{1}{2}\gamma \left(\frac{1+c-2c^{d+1}}{1-c} + (1+L)c^d \right) \|e_k^0\|. \end{aligned} \quad (51)$$

The above inequality actually provides an upper bound on the worst-case deterioration of the approximation to the Jacobian after d iterates and a switch from the k th to the $(k+1)$ st problem. There are two special cases of interest:

- 1) H is linear, hence $\gamma = 0$, and we then obtain uniform superlinear convergence of the sequence of problems $\{P_k(x_k) = 0\}$, which is essentially reduced to one single problem to which Broyden's method is applied. This yields an asymptotic observer different from the classical Luenberger observer.
- 2) The system is operated near an equilibrium point, in which case the term $\|x_{k+1} - x_k\|$ is small, or H is weakly nonlinear, in which case γ is small. In either case, $\|E_k^0\|$ will remain sufficiently small over a number of problems.

In general, due to the last two terms in (51), the sequence $\{\|A_k^0 - J(x_k)\|\}$ will not be uniformly bounded from above. Hence, for some k , A_k^0 will no longer be sufficiently close to $J(x^k)$, which, in practice, may be indicated by slow decrease, or even increase, of the term $\|Y_k - H(x_k^i)\|$, or by ill-conditioning of the matrix A_k^i . In an actual implementation of Broyden's method, one will, for reasons of computational complexity, typically update the QR -factorization of A_k^i , rather than A_k^i itself, and ill-conditioning will be checked for to avoid numerical instabilities [8]. For the iterates to converge throughout the sequence of problems, $J(x_k^0)$ will have to be recalculated occasionally. Although one might be able to obtain a tighter bound than (51), the term $\|x_{k+1} - x_k\|$ will remain. This term represents in a sense the "distance" between two subsequent problems, which is prescribed by the system's dynamics and, in general, cannot be made smaller such as to coerce uniform convergence of $\{x_k^0\}$, unless the observer is being used in a closed-loop situation and the state is being regulated to a fixed value, for example.

Suppose the discrete-time system (35) is the exact discretization with sampling time T of an underlying continuous-time system: $\dot{x} = f(x)$. Then the term $\|x_{k+1} - x_k\|$ in (51) can be estimated by

$$\|x_{k+1} - x_k\| = \|x((k+1)T) - x(kT)\| \leq \beta_f T \quad (52)$$

where β_f is the Lipschitz constant for f on some given set. Slower sampling means more time in between samples for calculating x_k given Y_k . However, (52) indicates also that the error in $\|A_{k+1}^0 - J(x_{k+1})\|$ grows faster as T increases, hence the Jacobian has to be recalculated more often too. On the other hand, as $T \rightarrow 0$, the approximation to the Jacobian will deteriorate more and more slowly. It should be pointed out though, that for very small T , the problem of solving $Y_k - H(x_k) = 0$ becomes ill-conditioned, since consecutive measurements will differ only slightly from each other. From a numerical analytic point of view, the system becomes practically unobservable. Thus, as is usual, there are trade-offs to be made.

VII. EXAMPLE: MIXED-CULTURE BIO REACTOR WITH COMPETITION AND EXTERNAL INHIBITION

In this section, we will illustrate the Newton observer and the continuous Newton observer by means of an example concerning a mixed-culture bioreactor. An application of the Broyden observer to an automotive problem can be found in [29].

The system under consideration describes the growth of two species in a continuously stirred bioreactor, which compete for a single rate-limiting substrate. In addition, an external agent is added which inhibits the growth of one species, while being deactivated by the second species. The measured quantity is the total cell mass of the two species. The example is taken from [18], where the system was shown to be globally feedback linearizable, provided that full state information is available.

Here, it will be shown first that, assuming noise-free measurements and dynamics, a discrete Newton observer may fail to converge if not properly initialized. The (global) continuous Newton observer is then implemented and shown to converge for a wide range of operating conditions irrespective of the observer initialization, i.e., even when large initial observer errors are present.

The system dynamics are given by

$$\begin{aligned} \dot{x}_1 &= \frac{0.4S(x_1, x_2)}{0.05 + S(x_1, x_2)} x_1 - u_1 x_1 \\ \dot{x}_2 &= \frac{0.01S(x_1, x_2)}{(0.05 + S(x_1, x_2))(0.02 + x_3)} x_2 - u_1 x_2 \\ \dot{x}_3 &= -0.5x_1 x_3 - u_1 x_3 + u_1 u_2 \end{aligned} \quad (53)$$

where $S(x_1, x_2) = 2 - 5x_1 - 6.667x_2$, time is expressed in hours.

- x_1 : cell density of inhibitor resistant species
- x_2 : cell density of inhibitor sensitive species
- x_3 : inhibitor concentration in fermentation medium
- u_1 : dilution rate
- u_2 : inlet concentration of the inhibitor.

Following the notation of the previous section, let $x_k := (x_k^1, x_k^2, x_k^3)^T$, $u_k := (u_k^1, u_k^2)^T$, and y_k denote the states, inputs, and outputs, respectively, evaluated at time kT , with T being the sampling time with which the system (53) is discretized. Furthermore, let $Y_k := (y_{k-2}, y_{k-1}, y_k)^T$ and $U_k := (u_{k-1}^1, u_k^1)^T$.

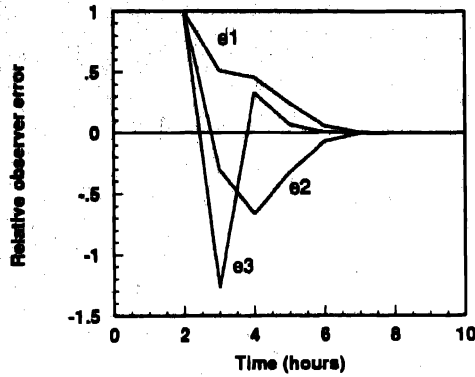


Fig. 1. Discrete Newton observer: relative observer error $e = (e_1, e_2, e_3)$, where $e_i = \frac{x_i - \hat{x}_i}{x_i}$, for $x = (0.2, 0.02, 0.005)$ and $\hat{x} = (0.02, 0.2, 0.015)$ when $u_1(t) = 0.3$ and $u_2(t) = 0.0067$.

A discretization of the above system with sampling time T may then be expressed as

$$\begin{aligned} x_{k+1} &= F_T^{u_k}(x_k) \\ y_k &= h(x_k) \end{aligned} \quad (54)$$

and the state-to-measurement map is given by

$$H(x_{k-2}, U_{k-1}) := \begin{pmatrix} h(x_{k-2}) \\ h \circ F^{u_{k-2}}(x_{k-2}) \\ h \circ F^{u_{k-1}} \circ F^{u_{k-2}}(x_{k-2}) \end{pmatrix} = Y_k. \quad (55)$$

Computing the rank of $\frac{\partial H}{\partial x}$ at a number of different points in the state space showed that the N -observability rank condition is indeed satisfied for $N = 3$. It was also found, however, that system (54) is poorly observable, indicated by ill-conditioning of $\frac{\partial H}{\partial x}$: ratio of its largest to smallest singular value for a sampling time of $T = 1$ (i.e., one hour) is on the order of 500. A typical response for the Newton observer is shown in Fig. 1, simulations showed however, that the Newton observer fails to converge if it is not initialized closely enough to the actual states, e.g., when

$$x = \begin{pmatrix} 0.2 \\ 0.02 \\ 0.005 \end{pmatrix} \text{ and } \hat{x} = \begin{pmatrix} 0.02 \\ 0.2 \\ 0.015 \end{pmatrix}.$$

Given the time scale—sampling times in the order of minutes or even hours—there is virtually no restriction to available CPU time in the observer design. We therefore simulated the continuous Newton observer as well. It is given by

$$\begin{aligned} \dot{z}_k^- &= K \left(\frac{\partial H}{\partial z}(z_k^-, U_{k-1}) \right)^{-1} [Y_k - H(z_k^-, U_{k-1})], \\ z_k^- &= F^{u_{k-2}}(z_k^-(kT)) \\ \hat{x}_k &= F^{u_{k-1}}(z_k^-). \end{aligned}$$

As expected, this observer did converge for all physically feasible initial values. Responses are shown in Fig. 2 for four values of the observer gain K . The plot confirms our finding that the observer may fail to converge if K is too small.

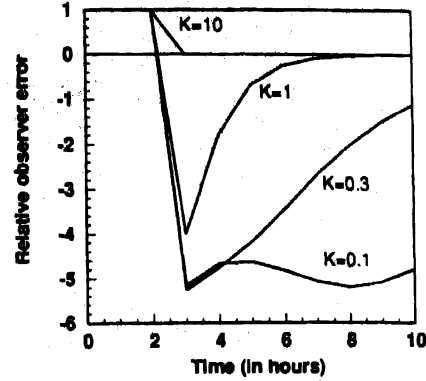


Fig. 2. Continuous Newton observer with different observer gains K : shown is the relative observer error for x_2 , when $u_1(t) \equiv 0.3$ and $u_2(t) \equiv 0.0067$.

VIII. CONCLUDING REMARKS

In this paper, we have provided a new observer design method for nonlinear systems with discrete measurements. The method relies on asymptotically inverting the state-to-measurement map, which is constructed by relating the system's state at a given time to a (predetermined) number of consecutive measurements. By using a continuous Newton method for the map inversion, the observer error was shown to converge to zero, globally and exponentially. If, instead, a computationally less expensive discrete Newton method is used, the observer shows quasilocal exponential convergence. Even more computational advantage may be gained by employing Broyden's method, whose effect on the observer performance was also investigated. The theory was illustrated on an example. Results on using this observer in a closed-loop setting have been reported in [17].

ACKNOWLEDGMENT

The authors wish to thank A. Tornambè for his insightful comments on the continuous Newton method.

REFERENCES

- [1] D. Aeyels, "Generic observability of differentiable systems," *Siam. J. Contr. Optim.*, vol. 19, pp. 595–603, 1981.
- [2] —, "On the number of samples necessary to achieve observability," *Syst. Contr. Lett.*, vol. 1, pp. 92–94, 1981.
- [3] B. M. Bell and F. W. Cathey, "The iterated Kalman filter update as a Gauss-Newton method," *IEEE Trans. Automat. Contr.*, vol. 38, no. 2, pp. 294–298, 1993.
- [4] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*. New York: Academic, 1975.
- [5] S. T. Chung, "Digital aspects of nonlinear synthesis problems," Ph.D. dissertation, University of Michigan, 1990.
- [6] S.-T. Chung and J. W. Grizzle, "Observer error linearization for sampled-data systems," *Automatica*, vol. 26, no. 6, pp. 997–1007, 1990.
- [7] W. F. Denham and S. Pines, "Sequential estimation when measurement function nonlinearity is comparable to measurement error," *AIAA J.*, vol. 4, pp. 1071–1076, 1966.
- [8] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
- [9] F. Deza, E. Busvelle, J. P. Gauthier, and D. Rakotopara, "High gain estimation for nonlinear systems," *Syst. Contr. Lett.*, vol. 18, pp. 295–299, 1992.
- [10] J. M. Fitts, "On the observability of nonlinear systems with applications to nonlinear regression analysis," *Inform. Sci.*, vol. 4, pp. 129–156, 1972.

- [11] B. A. Francis and T. T. Georgiou, "Stability theory for linear time-invariant plants with periodic digital controllers," *IEEE Trans Automat Contr*, vol. 33, no. 9, pp. 820-832, Sept 1988.
- [12] J. P. Gauthier, H. Hammouri, and S. Othman, "A simple observer for nonlinear systems, applications to bioreactors," *IEEE Trans Automat Contr*, vol. 37, pp. 875-880, June 1992.
- [13] S. T. Glad, "Observability and nonlinear deadbeat observers," in *Proc IEEE Conf Decis Contr*, San Antonio, TX, Dec 1983, pp. 800-802.
- [14] J. W. Grizzle and P. V. Kokotović, "Feedback linearization of sampled-data systems," *IEEE Trans Automat Contr*, vol. 33, pp. 857-859, Sep 1988.
- [15] J. W. Grizzle and P. E. Moraal, "Observer based control of nonlinear discrete-time systems," Univ of Michigan at Ann Arbor, Tech Rep CGR-39, College of Engineering, Control Group Reports, Feb 1990.
- [16] ———, *On Observers for Smooth Nonlinear Digital Systems* (Lecture Notes in Control and Information Sciences), vol. 144. Berlin: Springer-Verlag, 1990, pp. 401-410.
- [17] ———, "Newton, observers and nonlinear discrete-time control," in *Proc 29th CDC*, Hawaii, 1990, pp. 760-767.
- [18] K. A. Hoo and J. C. Kantor, "Global linearization and control of a mixed-culture bioreactor with competition and external inhibition," *Mathematical Biosciences*, vol. 82, pp. 43-62, 1986.
- [19] H. Th. Jongen, P. Jonker, and F. Twilt, *Optimization in Rⁿ*. Frankfurt am Main: Peter Lang Verlag, 1986.
- [20] S. Karahan, "Higher order linear approximations to nonlinear systems," Ph.D. dissertation, Mechanical Engineering, Univ of California Davis, 1989.
- [21] A. J. Krener, "Normal forms for linear and nonlinear systems," in *Differential Geometry: the Interface between Pure and Applied Mathematics*, M. Luksik, C. Martin, and W. Shadwick, Eds., vol. 68. Providence, RI: American Mathematical Society, 1986, pp. 157-189.
- [22] A. J. Krener and A. Isidori, "Linearization by output injection and nonlinear observers," *Syst Contr Lett*, vol. 3, pp. 47-52, 1983.
- [23] A. J. Krener, S. Karahan, M. Hubbard, and R. Frezza, "Higher order linear approximations to nonlinear control systems," in *Proc IEEE Conf Decis Contr*, LA, 1987, pp. 519-523.
- [24] A. J. Krener and W. Respondek, "Nonlinear observer with linearizable error dynamics," *SIAM J Contr*, vol. 47, pp. 1081-1100, 1988.
- [25] D. G. Luenberger, "Observers for multivariable systems," *IEEE Trans Automat Contr*, vol. AC-11, pp. 190-197, 1966.
- [26] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.
- [27] H. Michalska and D. Q. Mayne, "Moving horizon observers and observer based control," preprint, 1993.
- [28] P. E. Moraal and J. W. Grizzle, "Nonlinear discrete time observers using Newton's and Broyden's method," in *Proc ACC*, Chicago 1992, pp. 3086-3090.
- [29] P. E. Moraal, J. W. Grizzle, and J. A. Cook, "An observer design for single-sensor individual cylinder pressure control," to be presented at CDC, San Antonio, 1993.
- [30] K. T. Murty, *Linear Complementarity Linear and Nonlinear Programming*. Berlin: Heldermann Verlag, 1988.
- [31] S. Nicosia, A. Tornambè, and P. Valigi, "Use of observers for nonlinear map inversion," *Syst Contr Lett*, vol. 16, pp. 447-455, 1991.
- [32] H. Nijmeijer, "Observability of autonomous discrete-time nonlinear systems: a geometric approach," *Int J Contr*, vol. 36, pp. 867-874, 1982.
- [33] K. R. Shouse and D. G. Taylor, "Discrete-time observers for singularly perturbed continuous-time systems," to appear in *IEEE Trans Automat Contr*, 1993.
- [34] Y. Song and J. W. Grizzle, "The extended Kalman filter as a local asymptotic observer for nonlinear discrete-time systems," in *Proc 1992 ACC*, Chicago, 1992, pp. 3365-3369.
- [35] E. D. Sontag, "A concept of local observability," *Syst Contr Lett*, vol. 5, pp. 41-47, 1984.
- [36] A. Tornambè, "An asymptotic observer for solving the inverse kinematics problem," in *Proc Amer Contr Conf*, San Diego, May 1990, pp. 1774-1779.
- [37] J. Tsinias, "Further results on the observer design problem," *Syst Contr Lett*, vol. 14, pp. 411-418, 1990.
- [38] A. J. van der Schaft, "On nonlinear observers," *IEEE Trans Automat Contr*, vol. 30, no. 12, pp. 1254-1256, Dec 1985.
- [39] B. L. Walcott, M. J. Corless, and S. H. Zak, "Comparative study of nonlinear state-observation techniques," *Int J Contr*, vol. 45, no. 6, pp. 2109-2132, 1987.
- [40] P. J. Zufiria and R. S. Guttalu, "On an application of dynamical system theory to determine all the zeroes of a vector function," *J Math Analysis and Applic*, 152, pp. 269-295, 1990.



Paul E. Moraal was born in Drachten, The Netherlands in 1965. He received the engineering degree in applied mathematics from Twente University Enschede, The Netherlands in 1990 and the Ph.D. degree in electrical engineering systems from the University of Michigan Ann Arbor in 1994.

Dr. Moraal is currently an Engineering Specialist at Ford Motor Company's Research Laboratory in Dearborn, MI. His current research interests include nonlinear control theory and applications of modern control theory to automotive powertrain control.



Jessy W. Grizzle (S 79-M 83-SM 90) received the Ph.D. in electrical engineering from the University of Texas at Austin in 1983.

In 1984, he was at the Laboratoire des Signaux et Systèmes, Gif sur Yvette, France, as a PostDoc. From January 1984 to August 1987, he was an Assistant Professor in the Department of Electrical Engineering at the University of Illinois at Urbana-Champaign. Since September 1987, he has been with the Department of Electrical and Computer Science at the University of Michigan Ann Arbor.

where he is currently a Professor. He has held visiting appointments at the Dipartimento di Informatica e Sistemistica at the University of Rome and the Laboratoire des Signaux et Systèmes SUPELEC CNRS-SE Gif sur Yvette, France. He has served as a consultant on engine control systems to Ford Motor Company for seven years. His research interests include nonlinear system theory, geometric methods, multirate digital systems, automotive applications, and electronics manufacturing.

Dr. Grizzle was a Fulbright Grant awardee and a NATO Postdoctoral Fellow from January-December 1984. He received a Presidential Young Investigator Award in 1987 and the Henry Russell Award in 1993. In 1992, he received a Best Paper Award with K. L. Dobbins and J. A. Cook from the IEEE VEHICULAR TECHNOLOGY SOCIETY. In 1993, he received a College of Engineering teaching award. He was Past Associate Editor of TRANSACTIONS ON AUTOMATIC CONTROL and *System & Control Letters*.

A Probabilistic Approach to Multivariable Robust Filtering and Open-Loop Control

Kentth Öhrn, Anders Ahlén, *Senior Member, IEEE*, and Mikael Sternad, *Senior Member, IEEE*

Abstract—A new approach to robust filtering, prediction, and smoothing of discrete-time signal vectors is presented. Linear time-invariant filters are designed to be insensitive to spectral uncertainty in signal models. The goal is to obtain a simple design method, leading to filters which are not overly conservative. Modeling errors are described by sets of models, parameterized by random variables with known covariances. These covariances could either be estimated from data or be used as robustness “tuning knobs.” A robust design is obtained by minimizing the \mathcal{H}_2 -norm or, equivalently, the mean square estimation error, averaged with respect to the assumed model errors. A polynomial solution, based on an averaged spectral factorization and a unilateral Diophantine equation, is derived. The robust estimator is referred to as a cautious Wiener filter. It turns out to be only slightly more complicated to design than an ordinary Wiener filter. The methodology can be applied to any open-loop filtering or control problem. In particular, we illustrate this for the design of robust multivariable feedforward regulators, decoupling and model matching filters.

I. INTRODUCTION

FOR any model-based filter, modeling errors are a potential source of performance degradation. Here, we will propose a cautious Wiener filter for the prediction, filtering, or smoothing of discrete-time signal vectors. As in the scalar case, discussed in [36], it constitutes a generalization of the polynomial equations methodology pioneered by Kučera [21]. The design is based on a stochastic description of model errors, with relations to e.g., the stochastic embedding concept of Goodwin and coworkers [11], [12]. To be more specific, our problem formulation is as follows:

- A set of (true) dynamic systems is assumed to be well described by a set of discrete-time, stable, linear and time-invariant transfer function matrices

$$\mathcal{F} = \mathcal{F}_o + \Delta\mathcal{F}. \quad (1.1)$$

We call such a set an extended design model, in which \mathcal{F}_o represents a stable nominal model, while an error model $\Delta\mathcal{F}$ describes a set of stable transfer functions, parameterized by stochastic variables. The random variables enter linearly into $\Delta\mathcal{F}$, and they are assumed independent of the noise.

- A single robust linear filter is to be designed for the whole class of possible systems. Robust performance is obtained by minimizing the averaged mean square estimation error criterion

$$J = \text{trace} E E(\varepsilon(k) \varepsilon(k)^*). \quad (1.2)$$

Here, $\varepsilon(k)$ is the estimation error vector, E denotes expectation over noise and E is an expectation over the stochastic variables parameterizing the error model $\Delta\mathcal{F}$.

The averaged mean square error has been used previously in the literature by e.g., Chung and Bélanger [9], Speyer and Gustafson [32], and by Grimble [13]. These works were based on assumptions of small parametric uncertainties and on series expansions of uncertain parameters. We suggest the use of the criterion (1.2), together with a particular description of the set (1.1): transfer function elements in $\Delta\mathcal{F}$ have stochastic numerators and fixed denominators. Such models can describe nonparametric uncertainty and undermodeling as well as parametric uncertainty. A discussion of the utility, and versatility, of linearly parameterized stochastic error models can be found in [36].

Most previous suggestions for obtaining robust filters have been based on some type of minimax approach [10], [24]. A paper [26] by Martin and Mintz takes both spectral uncertainty and uncertainty in the noise distribution into account. The resulting filter will, however, be of very high order. Minimax design of a filter \mathcal{R} becomes very complex, unless there exists either a saddle point or a boundary point solution. A crucial condition here is that $\min_{\mathcal{R}} \max_{\mathcal{F}}$ equals $\max_{\mathcal{F}} \min_{\mathcal{R}}$. If so, one can search for models whose optimal filter gives the worst (nominal) performance and use the corresponding filter. As compared to finding the worst case with respect to a set of models, this is a much simpler task. It can still, however, be computationally demanding. See [19], [28], [31], [38], and the survey paper by Kassam and Poor [20]. The condition $\min_{\mathcal{R}} \max_{\mathcal{F}} = \max_{\mathcal{F}} \min_{\mathcal{R}}$ is not fulfilled in numerous problems, which makes them very difficult to solve. See, e.g., Example 5 in [36] and the example in Section IV.

Kalman filter-like estimators have recently been developed for systems with structured and possibly time-varying parametric uncertainty of the type

$$x(k+1) = (\mathbf{A} + \mathbf{D}\Delta(k)\mathbf{E})x(k) + w(k)$$

where the matrix $\Delta(k)$ contains norm-bounded uncertain parameters. See [30], [7], and [39] for continuous-time results and [40] for the discrete-time one-step predictor. See also [16]

Manuscript received July 2, 1993; revised May 20, 1994. Recommended by Associate Editor, B. Pasik-Duncan. The work was supported in part by Swedish Research Council for Engineering Sciences (TFR) under Grant 775.

The authors are with the Systems and Control Group, Uppsala University, Box 27, S-751 03 Uppsala, Sweden.

IEEE Log Number 9408271

for a related method. For systems which are stable for all $\Delta(k)$, an upper bound on the estimation error covariance matrix can be minimized by solving two coupled Riccati equations, combined with a one-dimensional numerical search. This represents a computational simplification, as compared to previous minimax designs. Still, the resulting estimators are quite conservative, partly because they rest on worst case design. This conservatism is illustrated and discussed in [29] and [37].

The method suggested in the present paper is computationally simpler than any of the minimax schemes referred to above. It also avoids two drawbacks of worst case designs. First, the stochastic variables in $\Delta\mathcal{F}$ need not have compact support. Thus, the descriptions of model uncertainties may have "soft" bounds. These are more readily obtainable in a noisy environment than the hard bounds required for minimax design. Second, not only the range of the uncertainties, but also their likelihood is taken into account by using the expectation $\bar{E}(\cdot)$ of the MSE. Highly probable model errors will affect the estimator design more than do very rare "worst cases." Therefore, the performance loss in the nominal case, the price paid for robustness, becomes smaller than for a minimax design. In other words, conservativeness is reduced. There do exist applications where a worst-case design is mandatory, e.g., for safety reasons. We believe, however, that the average performance of estimators is often a more appropriate measure of performance robustness.

In the present paper, one of our goals will be to present transparent design equations and to hold their number to a minimum without sacrificing numerical accuracy. We use matrix fraction descriptions with diagonal denominators and common denominator forms. This leads to a solution which is, in fact, significantly simpler and numerically better behaved than the corresponding nominal \mathcal{H}_2 -designs (without uncertainty) presented in [1] or [14]. Somewhat surprisingly, taking model uncertainty into account does not require any new types of design equations. We end up with just two equations for robust estimator design: a polynomial matrix spectral factorization and a unilateral Diophantine equation. The solution provides structural insight; important properties of a robust estimator are evident by direct inspection of the filter expression.

This paper is organized as follows. The filtering problem, model structure (1.1), and criterion (1.2) are discussed in more detail in Section II. Section III presents the design equations and some tools for performance evaluation. The design procedure is illustrated by a thorough numerical example in Section IV. The resulting estimator reduces the impact of model uncertainty and limited signal energy by using multiple sensors in an efficient way. In Section V the design of robust feedforward regulators, servos, and model matching filters is discussed.

Remarks on the Notation: Signals and polynomial coefficients may, in the following, be complex valued. (This is required in, e.g., communications applications.) Let p_j^* denote the complex conjugate (and transpose for matrices) of a polynomial coefficient p_j . For any polynomial

$$P(q^{-1}) = p_0 + p_1 q^{-1} + \dots + p_{np} q^{-np}$$

in the backward shift operator q^{-1} , define the conjugate polynomial

$$P_*(q) \triangleq p_0^* + p_1^* q + \dots + p_{np}^* q^{np}$$

where q is the forward shift operator. A polynomial $P(q, q^{-1})$ having coefficients of both q and q^{-1} will be called double-sided. Rational matrices, or transfer functions, are denoted by boldface calligraphic symbols, e.g., $\mathcal{R}(q^{-1})$. Polynomial matrices are denoted by boldface symbols, such as $\mathbf{P}(q^{-1})$, while constant matrices are denoted as \mathbf{P} . For example, the identity matrix of dimension n is denoted \mathbf{I}_n . We denote the trace of \mathbf{P} by $\text{tr} \mathbf{P}$. For polynomial or rational matrices, $P_*(q)$ and $\mathcal{R}_*(q)$ means complex conjugate, transpose, and substitution of q for q^{-1} . When appropriate, the complex variable z or $e^{j\omega}$ is substituted for the forward shift operator q . Arguments of polynomials and matrices are often omitted, when there is no risk of misunderstanding. The degree of a polynomial matrix is the highest degree of any of its polynomial elements. Square polynomial matrices $\mathbf{P}(q^{-1})$ are called stable if all zeros of $\det \mathbf{P}(z^{-1})$ are located in $|z| < 1$. A rational matrix is defined as stable if all its elements are stable. Causality is defined in the same way.

A rational matrix $\mathcal{G}(q^{-1})$ may be represented by polynomial matrices as a matrix fraction description (MFD), either left $\mathcal{G} = \mathbf{A}_1^{-1} \mathbf{B}_1$ or right $\mathcal{G} = \mathbf{B}_2 \mathbf{A}_2^{-1}$. It may also be represented in a common denominator form $\mathcal{G} = \mathbf{B}/\mathbf{A}$, where \mathbf{B} is a polynomial matrix. The scalar and monic polynomial A is then the least common denominator of all elements in \mathcal{G} . Denominator matrices in MFD's are assumed to have identity matrices as leading coefficients of their matrix polynomial representations, thus $\mathbf{A}_i(0) = \mathbf{I}$ above.

II. THE ROBUST ESTIMATION PROBLEM

Consider the following extended design model

$$\begin{aligned} y(k) &= \mathcal{G}(q^{-1})u(k) + \mathcal{H}(q^{-1})v(k) \\ u(k) &= \mathcal{F}(q^{-1})e(k) \\ f(k) &= \mathcal{D}(q^{-1})u(k) \end{aligned} \quad (2.1)$$

where \mathcal{G} , \mathcal{H} , \mathcal{F} , and \mathcal{D} are stable and causal, but possibly uncertain, transfer functions of dimension $p|s$, $p|r$, $s|n$, and $\ell|s$, respectively. The noise sequences $\{e(k)\}$ and $\{v(k)\}$ are mutually uncorrelated and zero mean stochastic sequences. To obtain a simple notation they are assumed to have unit covariance matrices, so scaling and uncertainty of the covariance are included in \mathcal{F} and \mathcal{H} , respectively. The signal $y(k)$ is assumed measurable, while $f(k)$ is the signal to be estimated

A. Multisignal Estimation

From data $y(k)$ up to time $k + m$, an estimator

$$\hat{f}(k|k+m) = \mathcal{R}(q^{-1})y(k+m) \quad (2.2)$$

of $f(k)$ is sought. See Fig. 1. The estimator may be a predictor ($m < 0$), a filter ($m = 0$), or a fixed lag smoother ($m > 0$). Here \mathcal{R} , of dimension $\ell|p$, is required to be stable and causal.

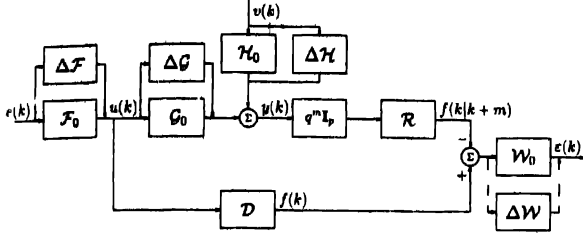


Fig. 1 A general linear filtering problem formulation. Based on noisy measurements $y(k+m)$, the signal $f(k)$ is to be estimated. Model errors in transfer functions are described by stochastic error models

The transfer function \mathcal{R} is designed to minimize the averaged mean square error (MSE) criterion (1.2)

$$J = \text{tr} \bar{E} E(\varepsilon(k) \varepsilon(k)^*) = \bar{E} E(\varepsilon(k)^* \varepsilon(k)) = \sum_{i=1}^l \bar{E} E|\varepsilon_i(k)|^2 \quad (2.3)$$

where

$$\varepsilon(k) = (\varepsilon_1(k) \cdots \varepsilon_l(k))^T \triangleq \mathcal{W}(q^{-1})(f(k) - \hat{f}(k|m)).$$

Above, \mathcal{W} is a stable and causal ℓ/ℓ rational weighting matrix, with a stable and causal inverse. It may be used by the designer to emphasize filtering performance in particular frequency bands. In filtering problems, \mathcal{W} is not assumed uncertain.

Model (2.1) offers considerable flexibility. For example, when estimating a signal $u(k)$ in colored noise, we set $\mathcal{G} = \mathcal{D} = \mathbf{I}_n$, giving $f(k) = u(k)$. In deconvolution, or input estimation problems, \mathcal{G} is a dynamic system and $\mathcal{D} = \mathbf{I}_n$. In a state estimation problem, $u(k)$ is the state vector, \mathcal{G} and \mathcal{D} are constant matrices while $\mathcal{H}v(k)$ represents (colored) measurement noise. Other special cases are discussed in [2], [3], and [8].

Example. An application where uncertain dynamics in \mathcal{G} is of interest is equalizer design for digital mobile radio communications [23]. A signal $u(k)$ then propagates along multiple paths, with different time delays, represented by delays in \mathcal{G} . The receiving antenna may have $p > 1$ elements (diversity design). See, e.g., [4]. Thus, an appropriate model of \mathcal{G} is a column vector of FIR channels, i.e., a vector of polynomials. The polynomials coefficients are estimated from short and noisy training sequences, with a known input $\{u(k)\}$. Estimation errors are inevitable. The task of a (robust) equalizer is to estimate $u(k)$, based on noisy measurements $y(k+m)$, a nominal model \mathcal{G}_o , and an estimate of the amount of model uncertainty \square

3 Parameterization of the Model

We choose to parameterize \mathcal{G} and \mathcal{H} as left MFD's having diagonal denominators,¹ while \mathcal{F} , \mathcal{D} , and \mathcal{W} are parameterized in common denominator form

$$\mathcal{G} = \mathbf{A}^{-1} \mathbf{B}; \quad \mathcal{H} = \mathbf{N}^{-1} \mathbf{M} \quad (2.4)$$

$$\mathcal{F} = \frac{1}{D} \mathbf{C}; \quad \mathcal{D} = \frac{1}{T} \mathbf{S}; \quad \mathcal{W} = \frac{1}{U} \mathbf{V}.$$

¹Note that this is a natural choice, if transfer functions are obtained by means of identification.

We have made these choices to obtain tidy and transparent design equations and to avoid coprime factorizations, which are known to be numerically sensitive. In (2.4), it is assumed that \mathcal{G} , \mathcal{H} , and \mathcal{F} may be uncertain. Introduction of uncertainty in the weighting matrix \mathcal{W} is not motivated in filtering problems. Its role in open-loop control will be discussed in Section III-B below. It is shown in Appendix C that uncertainty in \mathcal{D} does not affect the optimal filter design, provided it is uncorrelated to uncertainties in other blocks. Therefore, uncertainty in \mathcal{D} is not introduced.

The extended design models, cf. (1.1) and (2.1)

$$\mathcal{G} = \mathcal{G}_o + \Delta \mathcal{G}, \quad \mathcal{H} = \mathcal{H}_o + \Delta \mathcal{H}, \quad \mathcal{F} = \mathcal{F}_o + \Delta \mathcal{F}$$

are now expressed in polynomial matrix form. Using $\hat{\mathbf{B}}_o = \mathbf{A}_1 \mathbf{B}_o$, $\hat{\mathbf{B}}_1 = \mathbf{A}_o \mathbf{B}_1$ etc. we introduce

$$\begin{aligned} \mathcal{G} &= \mathbf{A}_o^{-1} \mathbf{B}_o + \mathbf{A}_1^{-1} \mathbf{B}_1 \Delta \mathbf{B} \\ &= \mathbf{A}_o^{-1} \mathbf{A}_1^{-1} (\hat{\mathbf{B}}_o + \hat{\mathbf{B}}_1 \Delta \mathbf{B}) \triangleq \mathbf{A}^{-1} \mathbf{B} \\ \mathcal{H} &= \mathbf{N}_o^{-1} \mathbf{M}_o + \mathbf{N}_1^{-1} \mathbf{M}_1 \Delta \mathbf{M} \\ &= \mathbf{N}_o^{-1} \mathbf{N}_1^{-1} (\hat{\mathbf{M}}_o + \hat{\mathbf{M}}_1 \Delta \mathbf{M}) \triangleq \mathbf{N}^{-1} \mathbf{M} \quad (2.5) \\ \mathcal{F} &= \frac{1}{D_o} \mathbf{C}_o + \frac{1}{D_1} \mathbf{C}_1 \Delta \mathbf{C} \\ &= \frac{1}{D_o D_1} (\hat{\mathbf{C}}_o + \hat{\mathbf{C}}_1 \Delta \mathbf{C}) \triangleq \frac{1}{D} \mathbf{C}. \end{aligned}$$

Above, $\mathcal{G}_o = \mathbf{A}_o^{-1} \mathbf{B}_o$ represents the nominal model and $\Delta \mathcal{G} = \mathbf{A}_1^{-1} \mathbf{B}_1 \Delta \mathbf{B}$ the error model. The same holds for \mathcal{H} and \mathcal{F} . The diagonal polynomial matrices $\mathbf{A} = \mathbf{A}_o \mathbf{A}_1$, $\mathbf{N} = \mathbf{N}_o \mathbf{N}_1$, and the polynomials $D = D_o D_1$, T and U are all assumed to be stable, with causal inverses. Denominator polynomials are assumed monic. In the error models, the polynomial D_1 , the diagonal matrices \mathbf{A}_1 and \mathbf{N}_1 , and the matrices \mathbf{C}_1 , \mathbf{B}_1 and \mathbf{M}_1 are fixed. They can be used to tailor the error models for specific needs. For example, if multiplicative error models are deemed appropriate, we use $\mathbf{A}_1 = \mathbf{A}_o$, $\mathbf{B}_1 = \mathbf{B}_o \mathbf{B}_m$ etc., with \mathbf{B}_m to be specified.

The matrices $\Delta \mathbf{B}$, $\Delta \mathbf{C}$, and $\Delta \mathbf{M}$ contain polynomials, with jointly distributed random variables as coefficients. These coefficients parameterize the class of assumed true systems. One particular modeling error is represented by one particular realization of the random coefficients.² Element ij of a stochastic polynomial matrix $\Delta \mathbf{P}$ is denoted

$$\Delta P^{ij} \triangleq [\Delta \mathbf{P}]_{ij} = \Delta p_o^{ij} + \Delta p_1^{ij} q^{-1} + \cdots + \Delta p_{h_p}^{ij} q^{-h_p} \quad (2.6)$$

where h_p is the degree of $\Delta \mathbf{P}$, i.e., the highest degree appearing in any polynomial ΔP^{ij} . All coefficients have zero means, so the nominal model is the average model in the set. Only the second-order moments of the random coefficients need to be specified, since the type of distribution, and higher order moments, will not affect the filter design. The parameter

²For a given system realization, the coefficients are assumed time-invariant and independent of the time-series $u(k)$ and $v(k)$. This is in contrast to the approach of Haddad and Bernstein in [15], who represent the effect of uncertainties by multiplicative noises. For a given uncertainty variance, a noise representation would underestimate the true effect of (time-invariant) parameter deviations in the dynamics.

covariances are denoted $\bar{E}(\Delta p_o^{ij})(\Delta p_o^{\ell k})^*$ and are collected in covariance matrices $\mathbf{P}_{\Delta P}^{(ij, \ell k)}$; see Section II-C.

We now introduce the assumption

- A1. The coefficients of all polynomial elements in ΔC are independent of those in ΔB .

It is possible to exclude Assumption A1, but it does simplify the solution, and it is also reasonable in most practical cases.

Error models can be obtained from ordinary identification experiments, provided the model structures match. For SISO systems, error models can be estimated in presence of under-modeling, using a maximum likelihood approach [11]. Even if the statistics is hard to obtain, one could still use the elements of covariance matrices pragmatically, as robustness "tuning knobs." They are then used similarly as when weighting matrices are adjusted in LQG controller design. An objective could be to obtain reasonable performance for the uncertainty set, for a prespecified acceptable degradation of performance in the nominal case. The error models may also be used to account for a slowly time-varying dynamics [25].

One way of obtaining the models (2.5)–(2.6) is by series expansion of state-space models with parametric uncertainty [37]. Parameter deviations are represented by stochastic variables. For small uncertainties, a first-order expansion can be used, which will directly lead to models of type (2.5). For larger uncertainties, a second-order Taylor expansion is usually sufficient; see [29]. Error models for nonparametric uncertainties can be adjusted directly to frequency domain data. In that context, a very useful concept is provided by the stochastic frequency domain theory of Goodwin and Salgado; see [12].

C. Covariance Matrices for the Stochastic Coefficients

To represent the uncertainties of the system in a natural way, covariance matrices will be organized as follows. The ij th element of a stochastic polynomial matrix ΔP can be expressed as

$$\Delta P^{ij}(q^{-1}) = \varphi^T(q^{-1}) \bar{p}_{ij} \quad (2.7)$$

where

$$\varphi^T(q^{-1}) = (1 \ q^{-1} \ \dots \ q^{-\delta p}) ; \quad \bar{p}_{ij} = (\Delta p_o^{ij} \Delta p_1^{ij} \ \dots \ \Delta p_{\delta p}^{ij})^T. \quad (2.8)$$

The cross covariance matrix $\mathbf{P}_{\Delta P}^{(ij, \ell k)}$, of dimension $\delta p + 1 | \delta p + 1$, between coefficients of $\Delta P^{ij}(q^{-1})$ and $\Delta P^{\ell k}(q^{-1})$, is given by

$$\begin{aligned} \mathbf{P}_{\Delta P}^{(ij, \ell k)} &= \bar{E} \bar{p}_{ij} \bar{p}_{\ell k}^* \\ &= \begin{bmatrix} \bar{E}(\Delta p_o^{ij})(\Delta p_o^{\ell k})^* & \dots & \bar{E}(\Delta p_o^{ij})(\Delta p_{\delta p}^{\ell k})^* \\ \vdots & \ddots & \vdots \\ \bar{E}(\Delta p_{\delta p}^{ij})(\Delta p_o^{\ell k})^* & \dots & \bar{E}(\Delta p_{\delta p}^{ij})(\Delta p_{\delta p}^{\ell k})^* \end{bmatrix} \end{aligned} \quad (2.9)$$

where $\mathbf{P}_{\Delta P}^{(ij, ij)}$ is Hermitian and positive semidefinite, while $\mathbf{P}_{\Delta P}^{(ij, \ell k)} = (\mathbf{P}_{\Delta P}^{(\ell k, ij)})^*$. Thus

$$\bar{E}(\Delta P^{ij} \Delta P^{\ell k}) = \bar{E}(\varphi^T(q^{-1}) \bar{p}_{ij} \bar{p}_{\ell k}^* \varphi(q)) = \varphi^T \mathbf{P}_{\Delta P}^{(ij, \ell k)} \varphi \quad (2.10)$$

With autocovariances, $(ij) = (\ell k)$, we model the uncertainty within each input–output pair. Cross-dependencies between different transfer functions may also be known. For example, uncertainty in one single physical parameter may very well enter into several transfer functions between inputs and outputs. Such effects are captured by cross covariances, $(ij) \neq (\ell k)$.

We collect all matrices of type (2.9) into one large covariance matrix, organized as shown by (2.11) at the bottom of the page. If ΔP has dimension $n|m$, then $\mathbf{P}_{\Delta P}$ is composed of nm by nm covariance matrices $\mathbf{P}_{\Delta P}^{(ij, \ell k)}$. The structure of (2.11) is useful from a design point of view. If, for example, a multivariable moving average model, or FIR model, is to be identified, then (2.11) is the natural way of representing the covariance matrix. If we instead prefer to use the blocks $\mathbf{P}_{\Delta P}^{(ij, \ell k)}$ of (2.11) as multivariable "tuning knobs," a given amount of uncertainty can be assigned to a specific input–output pair

III. DESIGN OF ROBUST FILTERS

A. An Averaged Spectral Factorization

We define an averaged spectral factor $\beta(q^{-1})$ as the numerator polynomial matrix of an averaged innovations model. It constitutes a key element of the robust filter. The average over the set of models, of the spectral density matrix $\Phi_y(e^{j\omega})$ of the measurement $y(k)$ is given by

$$\bar{E}\{\Phi_y(e^{j\omega})\} = \frac{1}{DD_*} A^{-1} N^{-1} \beta \beta_* N_*^{-1} A_*^{-1}.$$

$$\mathbf{P}_{\Delta P} = \begin{pmatrix} \begin{bmatrix} \mathbf{P}_{\Delta P}^{(11,11)} & \dots & \mathbf{P}_{\Delta P}^{(11,1m)} \\ \vdots & \ddots & \vdots \\ \mathbf{P}_{\Delta P}^{(1m,11)} & \dots & \mathbf{P}_{\Delta P}^{(1m,1m)} \end{bmatrix} & \dots & \begin{bmatrix} \mathbf{P}_{\Delta P}^{(11,n1)} & \dots & \mathbf{P}_{\Delta P}^{(11,nm)} \\ \vdots & \ddots & \vdots \\ \mathbf{P}_{\Delta P}^{(1m,n1)} & \dots & \mathbf{P}_{\Delta P}^{(1m,nm)} \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} \mathbf{P}_{\Delta P}^{(n1,11)} & \dots & \mathbf{P}_{\Delta P}^{(n1,1m)} \\ \vdots & \ddots & \vdots \\ \mathbf{P}_{\Delta P}^{(nm,11)} & \dots & \mathbf{P}_{\Delta P}^{(nm,1m)} \end{bmatrix} & \dots & \begin{bmatrix} \mathbf{P}_{\Delta P}^{(n1,n1)} & \dots & \mathbf{P}_{\Delta P}^{(n1,nm)} \\ \vdots & \ddots & \vdots \\ \mathbf{P}_{\Delta P}^{(nm,n1)} & \dots & \mathbf{P}_{\Delta P}^{(nm,nm)} \end{bmatrix} \end{pmatrix} \quad (2.11)$$

square polynomial matrix $\beta(z^{-1})$ is given by the stable solution to

$$\beta\beta_* = E\{NBCC_*B_*N_* + DAMM_*A_*D_*\} \quad (3.1)$$

Since that N^{-1} and A^{-1} are diagonal and will thus commute with the averaged second-order statistics of $y(k)$ is thus described by the same spectral density as for a vector-ARMA model

$$y(k) = \frac{1}{D} A^{-1} N^{-1} \beta \epsilon(k) \quad (3.2)$$

where $\epsilon(k)$ is white with a unit covariance matrix. This model is denoted as the averaged innovations model. (Note that $y(k) \neq y(k)$, but $\Phi_y(\epsilon^{i\omega}) = E\{\Phi_y(\epsilon^{i\omega})\}$.) When constructing the right-hand side of (3.1), the following results are useful.

Lemma 1 Let $H(q, q^{-1})$ be an $m|m$ polynomial matrix with double-sided polynomial elements having stochastic coefficients. Also, let $G(q, q^{-1})$ be an $n|m$ polynomial matrix with polynomial elements having stochastic coefficients, independent of all those of H . Then

$$E[GHG_*] = E[GE(H)G_*] \quad (3.3)$$

□

Proof. See Appendix A.

Now, introduce the double-sided polynomial matrices

$$\begin{aligned} CC_* &\triangleq L(CC_*) & B_\epsilon B_{\epsilon*} &\triangleq E(BCC_*B_*), \\ MM_* &\triangleq E(MM_*) \end{aligned} \quad (3.4)$$

Invoking (2.5) and using the fact that the stochastic coefficients are assumed to be zero mean, gives

$$\begin{aligned} CC_* &= C_o C_{o*} + C_1 E(\Delta C \Delta C_*) C_{1*} \\ B_\epsilon B_{\epsilon*} &= B_o CC_* B_{o*} + B_1 E(\Delta B CC_* \Delta B_*) B_{1*} \\ MM_* &= M_o M_{o*} + M_1 E(\Delta M \Delta M_*) M_{1*} \end{aligned} \quad (3.5)$$

Factorizations to obtain C , B_ϵ , etc. need not be performed. The double-sided polynomial matrices are expressed as CC_* , etc. merely to simplify the notation.

Lemma 2 Let Assumption A1 hold. By using (3.4), (3.5) and invoking Lemma 1, the averaged spectral factorization (3.1) can be expressed as

$$\beta\beta_* = NB_\epsilon B_{\epsilon*} N_* + DAMM_* A_* D_* \quad (3.6)$$

□

Proof. See Appendix A.

With a given right-hand side, (3.6) is just an ordinary polynomial matrix left spectral factorization. It is solvable under the following mild assumption

- **A2.** The averaged spectral density matrix $E\{\Phi_y(\epsilon^{i\omega})\}$ is nonsingular for all ω .

This assumption is equivalent to the right-hand side of (3.6) being nonsingular on $|z| = 1$. Then, the solution to (3.6) is unique, up to a right unitary factor (If $HH_* = I$, then $\beta\beta_* = (\beta H)(H_*\beta_*)$.) Under Assumption A2, a solution exists, with β having nonsingular leading coefficient matrix

$\beta(0)$. Its degree, n/β , will be determined by the maximal degree of the two right-hand side terms of (3.6).¹

To obtain the right-hand side of (3.6), averaged polynomial matrices $\bar{E}(\Delta PH \Delta P_*)$ have to be computed, where $H(q, q^{-1}) = \bar{C}C_*$ or I . It is shown in Appendix B that the ij th element of $\bar{E}(\Delta PH \Delta P_*)$ is given by

$$E[\Delta PH \Delta P_*]_{ij} = \text{tr} H \begin{bmatrix} \varphi^I & 0 \\ 0 & \varphi^J \end{bmatrix} \times \begin{bmatrix} P_{\Delta P}^{(i1,j1)} & P_{\Delta P}^{(im,j1)} \\ P_{\Delta P}^{(i1,jm)} & P_{\Delta P}^{(im,jm)} \end{bmatrix} \begin{bmatrix} \varphi_*^T & 0 \\ 0 & \varphi_*^J \end{bmatrix} \quad (3.7)$$

where φ^T was defined in (2.8). The block covariance matrix in (3.7) constitutes the block-transpose of the ij th block $\begin{bmatrix} & \\ & \end{bmatrix}$ of $P_{\Delta P}$ in (2.11). Average factors in (3.5) are readily obtained by substituting ΔC , ΔB , and ΔM for ΔP in (3.7).

B A Second Spectral Factorization

In the feedforward control problems discussed in Section V, we shall allow for \mathcal{W} being uncertain. Using a common denominator form, \mathcal{W} is parameterized in a similar way as \mathcal{F} in (2.5))

$$\mathcal{W} = V_o \frac{1}{U_o} + \Delta V V_1 \frac{1}{U_1} = \frac{1}{U_o U_1} (V_o + \Delta V V_1) \triangleq \frac{1}{U} V \quad (3.8)$$

A stable square matrix V , with $V(0)$ nonsingular, is introduced as a solution of the right spectral factorization

$$V_* V = E(V_* V) = V_{o*} V_o + V_{1*} E(\Delta V_* \Delta V) V_1 \quad (3.9)$$

Also, introduce the following assumptions

- **A3.** The coefficients of ΔV are independent of all other stochastic coefficients.
- **A4.** The right-hand side of (3.9) is nonsingular on the unit circle.

Whenever \mathcal{W} is known, ($\mathcal{W} = V_o/U_o = V/U$), (3.9) need not be solved, and $V = V_o = V$. This will be the case in filtering problems.

C The Cautious Multivariable Wiener filter

Theorem 1 Assume an extended design model (2.1), (2.4), (2.5), (3.8) to be given, with known covariance matrices (2.11). Assume A1–A4 to hold. A realizable estimator of $f(k)$ then minimizes the averaged MSE (2.3), among all linear time-invariant estimators based on $y(k+m)$, if and only if it has the same coprime factors as

$$f(k|k+m) = \mathcal{R} y(k+m) = \frac{1}{T} V^{-1} Q \beta^{-1} N A y(k+m) \quad (3.10)$$

¹When solving (3.6), we have utilized an algorithm by Ježek and Kučera, presented in [18]. It provides a solution with an upper triangular full rank leading coefficient matrix. For an overview of spectral factorization algorithms, see [22].

Here, $\beta(q^{-1})$ is obtained from (3.6), $\tilde{V}(q^{-1})$ from (3.9) while $Q(q^{-1})$ together with $L_*(q)$, both of dimensions $\ell|p$, is the unique solution to the unilateral Diophantine equation

$$q^{-m}\tilde{V}S\tilde{C}\tilde{C}_*\hat{B}_{o*}N_* = Q\beta_* + qL_*UTDI_p \quad (3.11)$$

with generic⁴ degrees

$$\begin{aligned} nQ &= \max(n\tilde{v} + n\pi + n\tilde{c} + m, nu + nt + nd - 1) \\ nL_* &= \max(n\tilde{c} + n\hat{b}_o + nn - m, n\beta) - 1 \end{aligned} \quad (3.12)$$

where $n\pi = \deg S$ etc. When applying the estimator (3.10) on an ensemble of systems, the minimal criterion value becomes

$$\begin{aligned} \text{tr} \tilde{E}E(\varepsilon(k)\varepsilon(k)^*)_{\min} &= \frac{1}{2\pi j} \oint_{|z|=1} \left\{ L_*\beta_*^{-1}\beta^{-1}L + \frac{1}{UTDD_*T_*U_*} \right. \\ &\quad \times \tilde{V}S\tilde{C} \left[I_n - \tilde{C}_*\hat{B}_{o*}N_*\beta_*^{-1}\beta^{-1}N\hat{B}_o\tilde{C} \right] \\ &\quad \left. \times \tilde{C}_*S_*\tilde{V}_* \right\} \frac{dz}{z}. \end{aligned} \quad (3.13)$$

□

Proof: See Appendix C.

Remarks: The only new type of computation, as compared to the nominal case described in [1]–[3], is the calculation of averaged polynomials using (3.7).

Since both \tilde{V} and β are stable, the estimator \mathcal{R} will be stable.⁵ If Assumptions A2 and A4 hold, $\tilde{V}(0)$ and $\beta(0)$ are nonsingular, so \mathcal{R} will be causal.

Note that the diagonal matrix $NA = N_oN_1A_oA_1$ appears explicitly in the filter (3.10). Thus, important properties of the robust estimator are evident by direct inspection. For example, assume some diagonal elements of N_1^{-1} or A_1^{-1} in the error models to have resonance peaks, indicating large uncertainty at the corresponding frequencies. Then, the filter will have notches, so the filter gain from the uncertain components of $y(k+m)$ will be low at the relevant frequencies.

The nominal Wiener filter has as a component a whitening filter. The robust estimator has a similar structure. By multiplying \mathcal{R} by the stable common factor D/D , the filter in (3.10) will contain $\beta^{-1}NAD$ as right factor. This averaged counterpart of a whitening filter is the inverse of the averaged innovations model (3.2).

The model structure (2.4)–(2.5) was selected to obtain a few simple design equations. Other choices are possible, but lead to various complications. For example, if stochastic polynomials had been introduced in the denominators, no exact analytical solution could have been obtained. Also stability would have been a problem. The use of general left MFD representations, instead of forms with diagonal denominators or common denominators, would have led to a solution involving seven coprime factorizations. Such a solution is presented in [29], but it provides less physical insight. It does also exhibit worse numerical behavior, since algorithms for coprime factorization are numerically sensitive.

⁴In special cases, the degrees may be lower.

⁵Stable common factors may exist in (3.10). They could be detected by calculating invariant polynomials of the involved matrices. If such factors have zeros close to the unit circle, it is advisable to cancel them before the filter is implemented. Otherwise, slowly decaying (initial) transients may deteriorate the filtering performance.

Furthermore, (3.11) is a unilateral Diophantine equation since Q and L_* appear on the same side of the terms in which they are involved. (When the unknowns appear on opposite sides, the equation is bilateral.) This property is a consequence of our choice of U, T , and D as scalar polynomials. Unilateral equations can easily be transformed into a system of linear equations, $AX = B$, where A is a block-Toeplitz matrix. For an example, see Section IV.

Robust design also makes the solution less numerically sensitive. Almost common factors of $\det \beta_*$ and UTD with zeros close to $|z| = 1$ would make the solution of (3.11) numerically sensitive. In the presence of model uncertainty, the risk for this is less than in the nominal case, due to the presence of averaged factors in (3.5). The averaged spectral factor β will, in general, have its zeros more distant from the unit circle than the nominal spectral factor, given by (3.20) below. This reduces the numerical difficulty of solving both (3.6) and (3.11).

For every cautious Wiener filter, there exists a system (without uncertainty) for which this estimator is the optimal Wiener filter (see [29]). It is therefore possible to represent model uncertainties by colored noises and then design a Wiener filter for the corresponding system. This correspondence provides a way of understanding the structure of the above design equations. We do not recommend such an equivalent noise-approach in the actual design, however, for two reasons:

- It is far from trivial to obtain a noise spectrum having similar effect on the filter design as do uncertainties in the block \mathcal{G} in Fig. 1. This is true in particular if the block \mathcal{F} is also uncertain, and if the problem is multivariable.
- It is advantageous from a design point of view to have separate tools which handle different aspects. Error models should represent the effect of modeling uncertainty; noise models should represent disturbances; criterion weighting functions should reflect the priorities of the user. A method which does not distinguish between these aspects will tend to confuse the designer.

The attainable performance improves monotonically with an increasing smoothing lag m . The following result gives the lower bound of the averaged estimation error. This bound can be approached pointwise in the frequency domain for $m < \infty$ by using a criterion filter \mathcal{W} with a high resonance peak.

Corollary 1: The limiting estimator for $m \rightarrow \infty$, the nonrealizable cautious Wiener filter, can be expressed as

$$\lim_{m \rightarrow \infty} q^m \mathcal{R} = \frac{1}{T} S\tilde{C}\tilde{C}_*\hat{B}_{o*}N_*\beta_*^{-1}\beta^{-1}NA. \quad (3.14)$$

Its average performance is given by (3.13) with $L = 0$. If $\mathcal{W} = I_\ell$, the spectral distribution of the lower bound of the estimation error $f(k) - \hat{f}(k|k+m)$ is

$$\begin{aligned} \lim_{m \rightarrow \infty} \text{tr} \tilde{E} \Phi_{f-f}(e^{j\omega}) &= \frac{1}{TDD_*T_*} \text{tr} \{ S\tilde{C} \\ &\quad [I_n - \tilde{C}_*\hat{B}_{o*}N_*\beta_*^{-1}\beta^{-1}N\hat{B}_o\tilde{C}] \tilde{C}_*S_* \}. \end{aligned} \quad (3.15)$$

The bound can be attained at a frequency ω_1 by an estimator with finite smoothing lag, if it is designed using a weighted criterion where

$$U(e^{-i\omega_1}) \approx 0. \quad (3.16)$$

□

Proof: In a similar way as in Appendix A.3 of [8], it is straightforward to show that $L \rightarrow 0$ as $m \rightarrow \infty$ in (3.11). Thus, (3.11) gives

$$\lim_{m \rightarrow \infty} q^m Q = \tilde{V} \tilde{S} \tilde{C} \tilde{C}^* \tilde{B}_{o*} N_* \beta_*^{-1}. \quad (3.17)$$

The substitution of this expression into (3.10) gives (3.14). The use of $L = 0$, $\tilde{V} = I_r$ and $U = 1$ in the integrand of (3.13) gives (3.15). When $U(e^{-i\omega_1}) \approx 0$, we obtain the same effect on the Diophantine equation (3.11) at the frequency ω_1 as if $L \rightarrow 0$: the rightmost term vanishes. Thus, at ω_1 , the gain and the phase of the elements of the polynomial matrix $q^m Q$ are approximately equal to those of (3.17) and the estimation error approaches the lower bound (3.15). ■

Remarks: Note that for realizable estimators (m finite), the lower bound (3.15) is only attainable at distinct frequencies ω_i by means of frequency weighting. For frequencies outside the bandwidth of \mathcal{W} , the estimate may be severely degraded. The results of Corollary 1 are illustrated at the end of the example in Section IV.

D. Analytical Expressions for Performance Evaluation

Theorem 2: Let a nominal estimator \mathcal{R}_n be designed based on a nominal model, with no uncertainties. Applying it, instead of (3.10), on an ensemble of systems results in an increase, as compared to (3.13), of the averaged MSE. The increase is given by

$$\text{tr} \tilde{E} E(\varepsilon(k) \varepsilon(k)^*) - \text{tr} \tilde{E} E(\varepsilon(k) \varepsilon(k)^*)_{\min} = \|\mathcal{W}(\mathcal{R}_n - \mathcal{R}) z^m D^{-1} A^{-1} N^{-1} \beta\|_2^2 \quad (3.18)$$

where β is defined by (3.1)–(3.6) and \mathcal{R} is the robust estimator (3.10). □

Proof: To obtain (3.18), the nominal filter \mathcal{R}_n is expressed as $\mathcal{R} + (\mathcal{R}_n - \mathcal{R})$. The optimality of \mathcal{R} implies that any modification gives an orthogonal contribution to the criterion. This, and the use of the averaged innovations model (3.2), gives (3.18). Mixed terms vanish, due to the orthogonality. ■

Theorem 3: Let a robust estimator \mathcal{R} be designed by (3.6)–(3.11). When applying it on a system equal to the nominal model, the increased MSE, as compared to the minimum obtainable with a nominal estimator \mathcal{R}_n , is

$$\text{tr} E(\varepsilon(k) \varepsilon(k)^*) - \text{tr} E(\varepsilon(k) \varepsilon(k)^*)_n = \|\mathcal{W}(\mathcal{R} - \mathcal{R}_n) z^m D_o^{-1} A_o^{-1} N_o^{-1} \beta_o\|_2^2. \quad (3.19)$$

Here, $D_o^{-1} A_o^{-1} N_o^{-1} \beta_o$ is the nominal innovations model and β_o is obtained from the nominal spectral factorization

$$\beta_o \beta_{o*} = N_o B_o C_o C_{o*} B_{o*} N_{o*} + D_o A_o M_o M_{o*} A_{o*} D_{o*}. \quad (3.20)$$

□

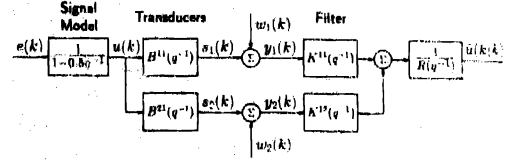


Fig. 2. Model and filter structure in the design example.

Proof: Analogous to that of Theorem 2, by expressing \mathcal{R} as $\mathcal{R}_n + (\mathcal{R} - \mathcal{R}_n)$. ■

Remarks: Expression (3.18) can be used for arbitrary linear estimators \mathcal{R}_n , for example, minimax-designs. Expression (3.19) quantifies the price paid in nominal performance for obtaining a robust design. The averaged innovations model in (3.18), and the nominal innovations model in (3.19), together with the filter \mathcal{W} , can be seen as weighting functions.

The largest effect of robust filtering is obtained at moderate and high signal-to-noise ratios. If the variance of broad-band measurement noise is increased, the gains of both the nominal and the robust filters decrease. If the noise level is high, performance differences between nominal and robust solutions tend to be small.

IV. A DESIGN EXAMPLE

Assume that a scalar signal $u(k)$ is to be estimated. It is described by a first order AR-process without uncertainty

$$u(k) = \frac{1}{1 - 0.5q^{-1}} c(k) \quad ; \quad E c(k)^2 = 1.$$

Thus, $\mathcal{D} = \mathcal{S}/T = 1$, $D_1 = 1$, $D = D_o = 1 - 0.5q^{-1}$, $\hat{C}_o = 1$, and $\hat{C}_1 = 0$. This signal is measured by two transducers ($p = 2$), with nominal models being second order FIR filters. The transducers are modeled by

$$y(k) = (B_o + A_1^{-1} \Delta B) u(k) + w(k)$$

with

$$B_o = \begin{pmatrix} B_o^{11} \\ B_o^{21} \end{pmatrix} = \begin{pmatrix} 0.100 + 0.080q^{-2} \\ 1 - 1.4q^{-1} + 0.92q^{-2} \end{pmatrix} \quad (4.1)$$

$$A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 - 0.6q^{-1} \end{pmatrix};$$

$$\Delta B = \begin{pmatrix} \Delta B^{11} \\ \Delta B^{21} \end{pmatrix} = \begin{pmatrix} \Delta b_o^{11} + \Delta b_2^{11} q^{-2} \\ \Delta b_o^{21} + \Delta b_1^{21} q^{-1} + \Delta b_2^{21} q^{-2} \end{pmatrix}.$$

Thus, $B_1 = A_o = I_2$ and $A = A_1$ are used in (2.5). See Fig. 2.

In the first transducer B^{11} , there is only a single uncertain parameter. It affects the coefficients Δb_o^{11} and Δb_2^{11} with opposite signs, so they have zero mean, variance r_1^2 and cross-covariance $-r_1^2$. In the second transducer, the stochastic coefficients are assumed mutually uncorrelated, with zero

means and equal variance r_2^2 . Thus, the auto-covariance matrices are

$$\mathbf{P}_{\Delta\mathbf{B}}^{(11\ 11)} = r_1^2 \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} \quad \mathbf{P}_{\Delta\mathbf{B}}^{(21\ 21)} = r_2^2 \mathbf{I}_3 \quad (4.2)$$

The scale factors (standard deviations) of the uncertainties are set to

$$r_1 = 0.02 \quad r_2 = 0.10 \quad (4.3)$$

Coefficients of ΔB^{11} and ΔB^{21} are assumed mutually uncorrelated. The complete covariance matrix (2.11) then becomes

$$\mathbf{P}_{\Delta\mathbf{B}} = \begin{pmatrix} \mathbf{P}_{\Delta\mathbf{B}}^{(11\ 11)} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{\Delta\mathbf{B}}^{(21\ 21)} \end{pmatrix} \quad (4.4)$$

The measurement noises $w_i(k)$ have variance 0.01. They are white and mutually uncorrelated. Thus, $w(k) = \mathbf{M}v(k)$, with

$$\mathbf{M} = \mathbf{M}_o = 0.1 \mathbf{I}_2 \quad (4.5)$$

The goal is now to design a filter ($m = 0$), which estimates $u(k)$ based on the two measurements $y_1(k)$ and $y_2(k)$. Frequency weighting is not used here ($\mathbf{W} = 1$), but its influence will be illustrated at the end of the example. In (3.4)–(3.5), we obtain

$$\mathbf{C}\mathbf{C}_* = E(\mathbf{C}\mathbf{C}_*) = 1$$

$$\mathbf{M}\mathbf{M}_* = E(\mathbf{M}\mathbf{M}_*) = 0.01 \mathbf{I}_2 \quad (4.6)$$

$$\mathbf{B}_c \mathbf{B}_{c*} = E(\mathbf{B}\mathbf{C}\mathbf{C}_* \mathbf{B}_*) = \mathbf{B}_o \mathbf{B}_{o*} + L(\Delta\mathbf{B}\Delta\mathbf{B}_*)$$

Expression (3.7) or (2.10) gives

$$\begin{aligned} E(\Delta\mathbf{B}^{11}\Delta\mathbf{B}_*^{11}) &= \varphi^T \mathbf{P}_{\Delta\mathbf{B}}^{(11\ 11)} \varphi_*^T \\ &= (1 - q^{-1}q^{-2})r_1^2 \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ q \\ q^2 \end{pmatrix} \\ &= r_1^2(-q^2 + 2 - q^{-2}) \end{aligned} \quad (4.7)$$

Note that $E(\Delta\mathbf{B}^{11}\Delta\mathbf{B}_*^{11})$ has zeros at $z = 1$ and at $z = -1$. Thus, the static gain and the high-frequency gain are assumed to be exactly known. Furthermore

$$E(\Delta\mathbf{B}^{21}\Delta\mathbf{B}_*^{21}) = \varphi^T \mathbf{P}_{\Delta\mathbf{B}}^{(21\ 21)} \varphi_*^T = 3r_2^2 \quad (4.8)$$

The spectral factorization (3.6) has dimension $p|p = 2|$. Using (4.6), it reduces to

$$\beta\beta_* = \mathbf{B}_c \mathbf{B}_{c*} + D\mathbf{A}\tilde{\mathbf{M}}\tilde{\mathbf{M}}_*\mathbf{A}_*D_*$$

$$= \mathbf{A}_1\mathbf{B}_o\mathbf{B}_{o*}\mathbf{A}_{1*} + E(\Delta\mathbf{B}\Delta\mathbf{B}_*) + 0.01D\mathbf{A}_1\mathbf{A}_{1*}D_*$$

where

$$E(\Delta\mathbf{B}\Delta\mathbf{B}_*) = \begin{pmatrix} r_1^2(-q^2 + 2 - q^{-2}) & 0 \\ 0 & 3r_2^2 \end{pmatrix}$$

By using the Newton-based algorithm described in [18], a stable averaged spectral factor, with $\beta(0)$ nonsingular, was found to be (4.9), as shown at the bottom of the page.

In the Diophantine equation (3.11), we use $m = 0$, $\mathbf{V} = \mathbf{V} = 1$, $\mathbf{S} = 1$, $\mathbf{C}\mathbf{C}_* = 1$, $\mathbf{N}_* = \mathbf{I}_2$, $\mathbf{U} = 1$, and $\mathbf{T} = 1$. Equation (3.11) thus reduces to

$$\mathbf{B}_{o*}\mathbf{A}_{1*} = \mathbf{Q}\beta_* + \mathbf{L}_*qD\mathbf{I}_2 \quad (4.10)$$

The degrees (3.12) are $n_Q = 0$, $n_{L_*} = 2$. By expressing the polynomial matrices as matrix polynomials, (4.10) becomes

$$\begin{aligned} (\mathbf{B}_0^* + \mathbf{B}_1^*q + \mathbf{B}_2^*q^2)(\mathbf{I}_2 + \mathbf{A}_1^*q) &= \mathbf{Q}_0(\beta_0^* + \beta_1^*q + \beta_2^*q^2 + \beta_3^*q^3) \\ &\quad + (\mathbf{L}_0^* + \mathbf{L}_1^*q + \mathbf{L}_2^*q^2) \\ &\quad \times (-0.5 + q)\mathbf{I}_2 \end{aligned}$$

Transpose this equation and note that since the coefficient matrices are real-valued, $\mathbf{P}_i^{*T} = \mathbf{P}_i$. By equating the two sides for each power of q separately, a linear system of eight equations, in block-Toeplitz form, is obtained

$$\begin{pmatrix} \mathbf{B}_0 \\ \mathbf{B}_1 + \mathbf{A}_1\mathbf{B}_0 \\ \mathbf{B}_2 + \mathbf{A}_1\mathbf{B}_1 \\ \mathbf{A}_1\mathbf{B}_2 \end{pmatrix} = \begin{pmatrix} \beta_0 & -0.5\mathbf{I}_2 & \mathbf{0} & \mathbf{0} \\ \beta_1 & \mathbf{I}_2 & -0.5\mathbf{I}_2 & \mathbf{0} \\ \beta_2 & \mathbf{0} & \mathbf{I}_2 & -0.5\mathbf{I}_2 \\ \beta_3 & \mathbf{0} & \mathbf{0} & \mathbf{I}_2 \end{pmatrix} \begin{pmatrix} \mathbf{Q}_0^T \\ \mathbf{L}_0^T \\ \mathbf{L}_1^T \\ \mathbf{L}_2^T \end{pmatrix}$$

With numerical values from (4.1) and (4.9), we obtain (x), as found at the bottom of the page. The solution is

$$\mathbf{Q} = (0.4005 \quad 0.7746)$$

$$\mathbf{L}_* = (0.0290 + 0.0200q \quad -0.2053 + 0.3224q - 0.1299q^2) \quad (4.11)$$

$$\beta = \begin{pmatrix} 0.1339 - 0.01867q^{-1} + 0.01622q^{-2} & 0.07862 - 0.01488q^{-1} + 0.06905q^{-2} \\ -0.1474q^{-1} + 0.2908q^{-2} - 0.1325q^{-3} & 1.1585 - 2.0327q^{-1} + 1.6219q^{-2} - 0.4765q^{-3} \end{pmatrix} \quad (4.9)$$

$$\begin{pmatrix} 1 \\ 1 \\ 0 \\ -2 \\ 0.8 \\ 1.76 \\ 0 \\ -5.52 \end{pmatrix} = \begin{pmatrix} 1.339 & 0.7862 & -5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.1585 & 0 & -5 & 0 & 0 & 0 & 0 \\ -0.1867 & -0.1488 & 1 & 0 & -5 & 0 & 0 & 0 \\ -1.474 & -2.0327 & 0 & 1 & 0 & -5 & 0 & 0 \\ 0.1622 & 0.6905 & 0 & 0 & 1 & 0 & -5 & 0 \\ 2.908 & 1.6219 & 0 & 0 & 0 & 1 & 0 & -5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ -1.325 & -4.765 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} Q_0^{11} \\ Q_0^{12} \\ \ell_0^{11} \\ \ell_0^{12} \\ \ell_1^{11} \\ \ell_1^{12} \\ \ell_2^{11} \\ \ell_2^{12} \end{pmatrix} \quad (x)$$

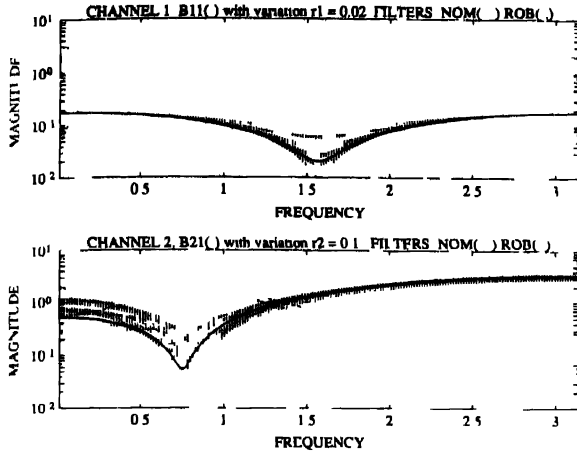


Fig 3 Bode magnitude plots for the nominal models of the two transducers $B^{11}(q^{-1})$ and $B^{21}(q^{-1})$ (solid). The dotted curves show fifteen realizations of possible true systems. Magnitude plots for the gains from $u_1(k)$ (upper) and $u_2(k)$ (lower) are shown for the robust estimator (dash dotted) and for the nominal Wiener filter design (dashed).

Finally the robust estimator (3.10) becomes

$$\mathbf{R} = \mathbf{Q}\beta^{-1}\mathbf{A}_1 = \frac{1}{R_r}(K_r^{11} K_r^{12}) \quad (4.12)$$

where the monic denominator is $R_r(q^{-1}) = \det \beta(q^{-1}) / \det \beta_0$. We obtain

$$\begin{aligned} K_r^{11} &= 2.9922 - 1.5138q^{-1} + 2.7365q^{-2} - 0.5687q^{-3} \\ K_r^{12} &= 0.4675 - 0.3341q^{-1} - 0.06115q^{-2} + 0.05841q^{-3} \\ R_r &= 1 - 1.8193q^{-1} + 1.6043q^{-2} - 0.6584q^{-3} \\ &\quad + 0.08179q^{-4} + 0.009182q^{-5} \end{aligned}$$

A corresponding nominal estimator (with no uncertainty assumed) is given by $(K_n^{11} K_n^{12})/R_n$ with

$$\begin{aligned} K_n^{11} &= 0.7419 - 1.0943q^{-1} + 0.3617q^{-2} \\ K_n^{12} &= 0.8792 - 0.3767q^{-1} - 0.03145q^{-2} \\ R_n &= 1 - 1.7786q^{-1} + 1.4269q^{-2} - 0.3938q^{-3} \end{aligned}$$

Fig 3 shows the Bode magnitude plots for the robust and nominal estimators. Also shown is the nominal transducer model and 15 randomly chosen systems. These were generated by using $\mathbf{B} = \mathbf{B}_0 + \mathbf{A}_1^{-1}\Delta\mathbf{B}$, with covariance matrix (4.4) and Gaussian distributions. The channel B^{11} has its uncertainty concentrated around the notch while B^{21} is uncertain mainly at low frequencies.

The gains of the nominal estimator (dashed curves) are determined exclusively by the nominal signal to noise ratios. The gains of the robust estimator (dash-dotted) are determined by the balance between noise levels and model uncertainties in the two channels. For example, the robust filter “knows” that channel 1 is well known, as compared to channel 2. Consequently, a higher gain is used from $y_1(k)$ as compared to the nominal case, and a lower gain from $y_2(k)$. The difference, as compared to nominal design, is largest at low frequencies. There, the dynamics of channel 1 is almost perfectly known, while channel 2 is very uncertain. The nominal filter gain in channel 2 is an approximate inverse of the nominal transducer. In contrast to the nominal filter, the robust filter has hardly

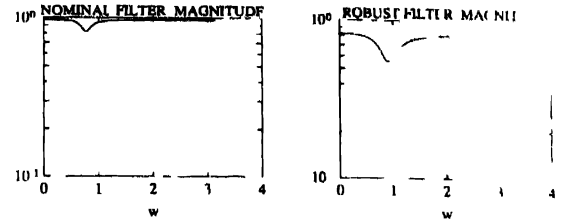


Fig 4 Bode magnitude plots for the transfer function from $u(k)$ to $u(k|k)$ for the nominal system with robust and nominal estimators.

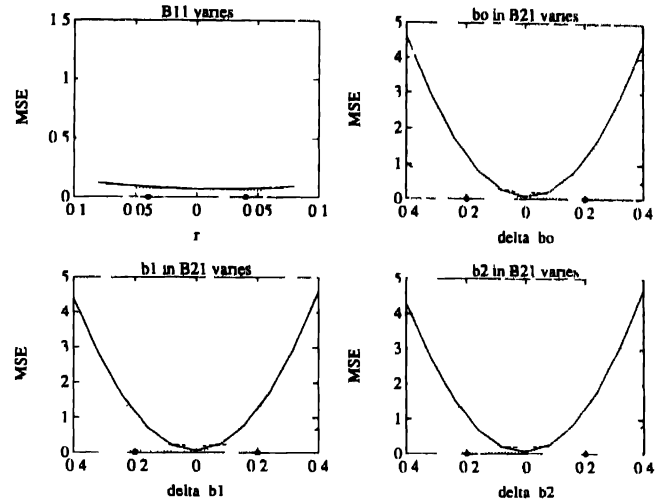


Fig 5 MSE for robust (dashed) and nominal filter (solid). One of the four uncertain parameters is varied while the others are held at nominal values. Also shown is the variance of $u(k)$ (upper dotted). It corresponds to the error caused by the trivial estimate $u(k) = 0$. The lower dotted curve is the lower bound achievable with knowledge of the true parameter values. Rings (o) indicate the two standard deviation limits of each parameter.

any peak at the (uncertain) notch around $\omega = 0.7$. It utilizes channel 1 more at this frequency.

Fig 4 shows Bode magnitude plots of transfer functions from $u(k)$ to $u(k|k)$. Since the noise levels are rather low the nominal estimator performs an almost complete inversion of the nominal transducers. The robust estimator is somewhat more cautious, but it also accomplishes a rather good inversion. It utilizes the two measurement signals differently than the nominal estimator.

Fig 5 shows the mean square estimation error when one of the uncertain parameters

$$\Delta b_1^{11} = -\Delta b_2^{11} \triangleq \tau, \quad \Delta b_0^{21} = \Delta b_1^{21} = \Delta b_2^{21}$$

is varied while the others are zero. The four parameters above span the set of assumed true systems, the extended design model. On average, over the four uncorrelated stochastic coefficients, the MSE is 0.32 for the robust filter and 0.90 for the nominal design. Note, however, that when τ is varied, the robust design (dashed) is actually slightly more sensitive than the nominal design. This is a price paid for reducing the sensitivity in the other dimensions.⁶

⁶It is natural that the robust filter has a somewhat increased sensitivity to model errors in channel 1: it has higher gain in that channel. This result is due to the much larger uncertainty in channel 2 at most frequencies. A designer worried about this effect could simply increase the value of the standard deviation τ_1 used in the design.

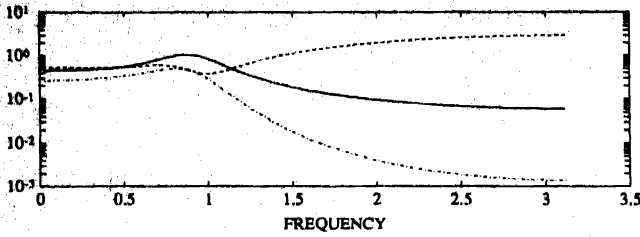


Fig. 6. The average spectral density of the estimation error $u(k) - \hat{u}(k|k)$. Shown in the plot are the lower bound, according to (3.15), (dash-dotted) and the average spectral densities obtained with (dashed) and without (solid) frequency weighting.

The robust estimator, of course, does not perform as well as the nominal one in the nominal case. This is mainly due to its somewhat lower gain from $u(k)$ to $\hat{u}(k|k)$; see Fig. 4. It is evident from Fig. 5 that this performance loss is very small, as compared to the improvement in nonideal situations.⁷

As an alternative, we tried to investigate minimax designs, i.e., worst case \mathcal{H}_2 -designs, assuming rectangular parameter distributions. This turned out to be prohibitively difficult, since no point where $\min_{\mathcal{R}} \max_{\Delta B} \varepsilon(k)^2 = \max_{\Delta B} \min_{\mathcal{R}} \varepsilon(k)^2$ could be found. These difficulties were in marked contrast to the ease of designing a cautious Wiener filter based on the averaged \mathcal{H}_2 -criterion.

Let us finally illustrate the effect of using a frequency-dependent weighting function in the criterion (2.3). Assume that the performance at frequencies close to $\omega = 0.9$ is of particular importance. The choice

$$\mathbf{W} = \frac{\mathbf{V}}{U} = \frac{1}{1 - 1.2184q^{-1} + 0.9604q^{-2}} \quad (4.13)$$

with a high resonance peak ($|z| = 0.98$) at $\omega = 0.9$ should, according to Corollary 1, result in a performance, at that frequency, close to the lower bound. Fig. 6 confirms this. Performance is substantially degraded at higher frequencies, however, where estimation accuracy is not emphasized.

V. ROBUST FEEDFORWARD CONTROL

A class of feedforward control problems turns out to be dual to the filtering problems discussed in Section II. We include a brief separate discussion of them, since it offers several engineering insights. Feedforward compensation does not affect the classical sensitivity function. The effect of an $x\%$ model deviation at a particular frequency, however, on e.g., the step response, will very much depend on the (nominal) magnitude of the transfer function at that frequency. As the gain at a particular frequency is increased by a feedforward link, model errors at that frequency become more and more noticeable. Therefore, it is of value to take model uncertainty into account explicitly in the feedforward design.

⁷It can also be noted that it is of advantage to use both channels. The minimal MSE, for channels equal to the nominal models, is 0.07 if both channels are used. It is 0.59 if only channel 1 is used and 0.11 if only channel 2 is used. The average MSE of the robust filter (0.32) is in fact lower than the nominal MSE for an estimator which uses only channel 1 (0.59).

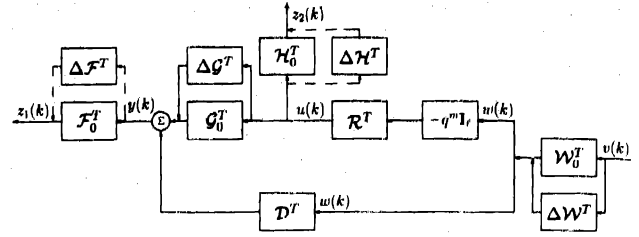


Fig. 7. A block diagram which is dual to the one in Fig. 1. Arrows are reversed, summation points and node points are interchanged, and transfer functions are transposed. The corresponding robust control problem is to design \mathcal{R}^T to minimize the average, over the class of models, of the \mathcal{H}_2 -norm of the transfer function from v to $z = (z_1^T \ z_2^T)^T$.

To stress the duality to filtering problems, the output of an uncertain but stable model will be described by

$$\begin{aligned} y(k) &= (\mathbf{G}_o^T + \Delta \mathbf{G}^T) u(k) + \mathbf{D}^T w(k) \\ &= (\mathbf{B}_o^T \mathbf{A}_o^{-T} + \Delta \mathbf{B}^T \mathbf{B}_1^T \mathbf{A}_1^{-T}) u(k) + \frac{1}{T} \mathbf{S}^T w(k) \end{aligned} \quad (5.1)$$

where \mathbf{A}^{-T} denotes inverse and transpose. The rational and polynomial matrices above have properties as outlined in Section II. The matrices \mathbf{G}^T and \mathbf{D}^T may contain delays. Based on possibly delayed or advanced measurements of

$$\begin{aligned} w(k) &= (\mathbf{W}_o^T + \Delta \mathbf{W}^T) v(k) \\ &= \left(\frac{1}{U_o} \mathbf{V}_o^T + \frac{1}{U_1} \Delta \mathbf{V}^T \mathbf{V}_1^T \right) v(k); \\ E(v(k)^T v(k)) &= \mathbf{I}_l \end{aligned} \quad (5.2)$$

a stable controller

$$u(k) = -\mathbf{R}^T w(k + m) \quad (5.3)$$

is to be designed to minimize the averaged \mathcal{H}_2 -norm of the transfer function from v to $(z_1^T \ z_2^T)^T \triangleq ((\mathcal{F}^T y)^T (\mathcal{H}^T u)^T)^T$

$$J' = E \left\| \begin{bmatrix} \mathcal{F}^T (\mathbf{D}^T - \mathbf{G}^T \mathbf{R}^T z^m) \mathbf{W}^T \\ -\mathcal{H}^T \mathbf{R}^T z^m \mathbf{W}^T \end{bmatrix} \right\|_2^2. \quad (5.4)$$

The control weighting $\mathcal{H}^T = \mathbf{M}^T \mathbf{N}^{-T}$ and the output weighting $\mathcal{F}^T = \mathbf{G}^T / D$ are normally specified by the designer and exactly known, i.e., $\Delta \mathcal{H}^T = 0$, $\Delta \mathcal{F}^T = 0$.

Problem formulation (5.1)–(5.4) may represent a disturbance measurement feedforward design. When $m > 0$, the disturbance $w(k)$ can be measured before it affects the system via \mathbf{D}^T . (Such a situation could, equivalently, be described by a delay q^{-m} in \mathbf{D}^T .) The formulation also covers reference feedforward problems, feedforward decoupling, and model matching. Then, $w(k)$ is a command signal and $\mathbf{W}^T v(k)$ is a (possibly uncertain) stochastic model, describing its second order properties. A servo filter \mathbf{R}^T is to be designed, so that the output $-\mathbf{G}^T u(k)$ optimally follows the response model $\mathbf{D}^T w(k)$.⁸ In decoupling problems, \mathbf{D}^T is diagonal.

The duality of feedforward control to the previously discussed filtering problem has been described in [5] for the

⁸The corresponding nominal result (without uncertainty) was discussed in [35] and in [5]. See also [27], where common denominator forms were used. The robust controller for SISO systems was presented in [36]. In [6], the applicability of the averaged \mathcal{H}_2 -criterion is demonstrated for a somewhat more general class of open-loop type problems.

nominal case. The corresponding result in the presence of model errors is given below.

Theorem 4: A feedforward filter, which solves the robust control problem (5.1)–(5.4) under assumptions A1–A4 is given by transposing \mathcal{R} from (3.10), or

$$u(k) = -A^T N^T \beta^{-T} Q^T \hat{V}^{-T} \frac{1}{T} w(k+m) \quad (5.5)$$

where β , Q , and \hat{V} are given by (3.6), (3.11), and (3.9), respectively. \square

Proof: The averaged \mathcal{H}_2 -norm is invariant under transposition. Thus, it fulfills the basic requirement of [5]. By extending the discussion in [5] to uncertain models the result is obtained. \blacksquare

The uncertainty ΔW^T of the disturbance or reference model (5.2) enters via the spectral factorization (3.9). Transposition of (3.9) gives

$$\hat{V}^T \hat{V}_*^T = U_1 V_o^T V_{o*}^T U_{1*} + U_o V_1^T \bar{E}(\Delta V^T \Delta V_*^T) V_{1*}^T U_{o*}. \quad (5.6)$$

This is the kind of left spectral factorization encountered when two noise sources are described by one innovations model. In fact, the uncertainty ΔW^T has exactly the same effect on the controller design as would a measurement noise on $w(k)$, with spectral density $V_1^T \bar{E}(\Delta V^T \Delta V_*^T) V_{1*}^T / U_1 U_{1*}$. We do not need to solve a right spectral factorization (3.9) in this problem. The left spectral factorization (5.6) can be solved instead.

As in the dual filtering case, uncertainty in the direct feedthrough, or response model $\mathcal{D}^T = S^T/T$ does not affect the optimal solution, if it is independent of the uncertainties in \mathcal{G}^T .

VI. CONCLUSIONS

A method for designing robust filters and feedforward controllers, based on imperfectly known linear models, has been presented. Modeling errors were described by sets of models, parameterized by random variables with known covariances. A robust design was obtained by minimizing the \mathcal{H}_2 -norm, averaged with respect to the assumed model errors. The estimator minimizes this criterion by balancing model uncertainties against noise properties, at different frequencies and in different measurement channels. When using robust filtering, the greatest sensitivity reduction is obtained at moderate and high signal to noise ratios. Dually, the largest impact of robust control is obtained for designs with low input penalties.

One variant of the discussed filtering problems is to explicitly define a part of the measurement vector as being a noise-free signal. This signal could e.g., represent known inputs to the system. Such a formulation is also of use in the optimization of decision feedback equalizers for digital communications [33], [34].

There exist efficient numerical algorithms based on a polynomial equations approach. (We have implemented them as MATLAB .m-files, and the code is available upon request.) For multivariable problems of high order and high signal vector dimension p , Riccati-based algorithms do, however, perform better numerically. For dimensions of the measurement vector

up to, say, four, algorithms based directly on polynomial manipulation can in general be used safely. For higher dimensions, we recommend analytical solutions to be obtained by an input–output approach, to gain engineering insight, but algorithms e.g., for spectral factorization to be based on state space formulations, cf. [22]

APPENDIX

A. Proofs of Lemmas

Proof of Lemma 1: Let $G = [G_1^T \cdots G_n^T]^T$, where $G_1 \cdots G_n$ represent the n polynomial row vectors of the $n|m$ matrix. Then, with H of dimension $m|m$, the ij th element of $\bar{E}(GHG_*)$ can be expressed as

$$\begin{aligned} [\bar{E}(GHG_*)]_{ij} &= \bar{E}(G_i H G_{j*}) \\ &= \bar{E}(\text{tr} G_{j*} G_i H) = \text{tr} \bar{E}(G_{j*} G_i) \bar{E}(H). \end{aligned}$$

In the last equality, we used the fact that all elements of $G_{j*} G_i$ are independent of all those of H . We also have that

$$\text{tr} \bar{E}(G_{j*} G_i) \bar{E}(H) = \bar{E}(G_i \bar{E}(H) G_{j*}) = [\bar{E}(G \bar{E}(H) G_*)]_{ij}$$

which proves (3.3), since $\bar{E}(\cdot)$ operates on all elements of GHG_* . \blacksquare

Proof of Lemma 2: If the coefficients of the elements of a polynomial matrix $\Delta P(q^{-1})$ are stochastic variables, then so are the coefficients of the elements in $\Delta P \Delta P_*$. The coefficients of the polynomial elements in ΔC and ΔB are independent and so are the coefficients of the elements in $\Delta C \Delta C_*$ and $\Delta B \Delta B_*$. By defining CC_* as H and ΔB as G in Lemma 1 and using (3.4), the right-hand side of (3.1) becomes

$$\begin{aligned} N \bar{E}[BCC_* B_*] N_* + D A \bar{E}(MM_*) A_* D_* &= \\ = N \bar{E}[B \bar{E}(CC_*) B_*] N_* + D A \bar{E}(MM_*) A_* D_* &= \\ = N \bar{E}[B \hat{C} \hat{C}_* B_*] N_* + D A \hat{M} \hat{M}_* A_* D_* & \end{aligned}$$

By once more utilizing (3.4), we obtain (3.6).

B. Calculation of Averaged Polynomial Matrices

Consider the matrices discussed in Lemma 1. Let the polynomial matrices $G = \Delta P$ and H be of dimensions $n|m$ and $m|m$, respectively. Denote the i -th row of ΔP by

$$\Delta P_i = [\Delta P^{i1} \cdots \Delta P^{im}]$$

where ΔP^{ij} are polynomials with stochastic coefficients. If H is assumed deterministic, the ij th element of $\bar{E}(\Delta P H \Delta P_*)$ can be written as

$$\begin{aligned} \bar{E}[\Delta P H \Delta P_*]_{ij} &= \text{tr} \bar{E}(\Delta P_{j*} \Delta P_i) H = \\ &= \text{tr} H \bar{E}[\Delta P_{j*}^{j1} \cdots \Delta P_{j*}^{jm}]^T [\Delta P^{i1} \cdots \Delta P^{im}] \\ &= \text{tr} H \bar{E} \begin{bmatrix} \Delta P^{i1} \Delta P_{j*}^{j1} & \cdots & \Delta P^{im} \Delta P_{j*}^{jm} \\ \vdots & \ddots & \vdots \\ \Delta P^{i1} \Delta P_{j*}^{jm} & \cdots & \Delta P^{im} \Delta P_{j*}^{jm} \end{bmatrix}. \end{aligned}$$

By using (2.10), we readily obtain (3.7). \blacksquare

C. Proof of Theorem 1

A technique for constructive derivation of polynomial design equations for Wiener filters was presented in [1]. This method is utilized here to minimize (2.3). The estimation error is given by

$$\varepsilon(k) = \mathbf{W}(f(k) - \hat{f}(k|k+m)). \quad (\text{C.1})$$

All admissible alternatives to a proposed estimate $\hat{f}(k|k+m)$, can be described by

$$\hat{d}(k) = \mathbf{R}y(k+m) + \hat{\nu}(k); \quad \hat{\nu}(k) = \mathbf{M}y(k+m). \quad (\text{C.2})$$

Here, \mathbf{M} is a rational, stable and causal, but otherwise arbitrary transfer function. Define the weighted variation

$$\begin{aligned} \nu(k) &\triangleq \mathbf{W}\hat{\nu}(k) \\ &= \frac{1}{U}\mathbf{V}\mathbf{M}q^m \left(\mathbf{A}^{-1}\mathbf{B}\frac{1}{D}\mathbf{C}e(k) + \mathbf{N}^{-1}\mathbf{M}v(k) \right). \end{aligned}$$

Optimality of (2.2) is obtained if no perturbation $\nu(k)$ will improve the average estimator performance. This occurs if and only if the error $\varepsilon(k)$ is orthogonal to any admissible weighted estimator variation $\nu(k)$. In other words

$$\text{tr} \bar{E}E(\varepsilon(k)\nu(k)^*) = \text{tr} \bar{E}E(\varepsilon(k)^*\nu(k)) = 0.$$

Then, the perturbed criterion value becomes

$$\begin{aligned} \bar{J} &= \text{tr} \bar{E}E[\mathbf{W}(f(k) - \hat{d}(k))][\mathbf{W}(f(k) - \hat{d}(k))]^* \\ &= \text{tr} \bar{E}E(\varepsilon(k)\varepsilon(k)^* - \varepsilon(k)\nu(k)^* \\ &\quad - \varepsilon(k)^*\nu(k) + \nu(k)\nu(k)^*) \\ &= \text{tr} \bar{E}E(\varepsilon(k)\varepsilon(k)^* + \nu(k)\nu(k)^*). \end{aligned} \quad (\text{C.3})$$

This expression is evidently minimized by $\nu(k) = 0$.

Since all transfer functions in (2.4) are assumed stable, both $\varepsilon(k)$ and $\nu(k)$ are stationary. Parseval's formula may then be used to express $\text{tr} \bar{E}E\varepsilon(k)\nu(k)^*$ as

$$\begin{aligned} &\text{tr} \bar{E}E \frac{1}{U} \mathbf{V} \left\{ \left(\frac{1}{T} \mathbf{S} - q^m \mathbf{R} \mathbf{A}^{-1} \mathbf{B} \right) \frac{1}{D} \mathbf{C} e(k) - q^m \mathbf{R} \mathbf{N}^{-1} \mathbf{M} v(k) \right\} \\ &\quad \times \left\{ \frac{1}{U} \mathbf{V} \mathbf{M} q^m \left(\mathbf{A}^{-1} \mathbf{B} \frac{1}{D} \mathbf{C} e(k) + \mathbf{N}^{-1} \mathbf{M} v(k) \right) \right\}^* \\ &= \text{tr} \bar{E} \frac{1}{2\pi j} \oint_{|z|=1} \frac{1}{UU_*DD_*} \\ &\quad \times \mathbf{V} \left\{ z^{-m} \frac{1}{T} \mathbf{S} \mathbf{C} \mathbf{C}_* \mathbf{B}_* \mathbf{A}_*^{-1} - \mathbf{R} \mathbf{A}^{-1} \mathbf{N}^{-1} \right. \\ &\quad \times (\mathbf{N} \mathbf{B} \mathbf{C} \mathbf{C}_* \mathbf{B}_* \mathbf{N}_* + \mathbf{D} \mathbf{A} \mathbf{M} \mathbf{M}_* \mathbf{A}_* \mathbf{D}_*) \mathbf{N}_*^{-1} \mathbf{A}_*^{-1} \} \\ &\quad \times \mathbf{M}_* \mathbf{V}_* \frac{dz}{z}. \end{aligned} \quad (\text{C.4})$$

Note that \mathbf{A} and \mathbf{N} commute, since they are diagonal. We are allowed to move the expectation \bar{E} inside the integration, since, for any particular realization of the elements of $\Delta \mathbf{C}$, $\Delta \mathbf{B}$, $\Delta \mathbf{M}$, and $\Delta \mathbf{V}$, all elements of the integrand are Riemann integrable on the unit circle; see e.g., [17, Theorem 3.8]. The use of the trace rotation, $\text{tr} \mathbf{V}\{\cdots\} \mathbf{M}_* \mathbf{V}_* = \text{tr} \mathbf{V}_* \mathbf{V}\{\cdots\} \mathbf{M}_*$,

the spectral factorizations (3.1) and (3.9) and the Assumption A3, leads to

$$\begin{aligned} \text{tr} \bar{E}E\varepsilon(k)\nu(k)^* &= \text{tr} \frac{1}{2\pi j} \oint \frac{1}{UU_*DD_*} \tilde{\mathbf{V}}_* \tilde{\mathbf{V}} \\ &\quad \times \left\{ z^{-m} \bar{E} \left(\frac{1}{T} \mathbf{S} \mathbf{C} \mathbf{C}_* \mathbf{B}_* \right) \mathbf{A}_*^{-1} \right. \\ &\quad \left. - \mathbf{R} \mathbf{A}^{-1} \mathbf{N}^{-1} \beta \beta_* \mathbf{N}_*^{-1} \mathbf{A}_*^{-1} \right\} \mathbf{M}_* \frac{dz}{z}. \end{aligned} \quad (\text{C.5})$$

From (C.5) it is now easy to see that uncertainties in

$$\mathbf{D} = \frac{1}{T_o} \mathbf{S}_o + \frac{1}{T_1} \mathbf{S}_1 \Delta \mathbf{S} = \frac{1}{T_o T_1} (\hat{\mathbf{S}}_o + \hat{\mathbf{S}}_1 \Delta \mathbf{S}) \triangleq \frac{1}{T} \mathbf{S} \quad (\text{C.6})$$

independent of $\Delta \mathbf{C}$, $\Delta \mathbf{B}$ and $\Delta \mathbf{V}$, will not affect the filter design since, using Assumption A1

$$\bar{E} \left(\frac{1}{T} \mathbf{S} \mathbf{C} \mathbf{C}_* \mathbf{B}_* \right) = \bar{E} \left(\frac{1}{T} \mathbf{S} \right) \bar{E} (\mathbf{C} \mathbf{C}_*) \bar{E} (\mathbf{B}_*) = \frac{1}{T_o} \mathbf{S}_o \tilde{\mathbf{C}} \tilde{\mathbf{C}}_* \hat{\mathbf{B}}_{o*}. \quad (\text{C.7})$$

(The minimal criterion value will be affected, however, which is evident from (C.12), below.) In the sequel we thus use $T_o = T$ and $\mathbf{S}_o = \mathbf{S}$.

Using (C.7) in (C.5) and extracting T to the left and \mathbf{A}_*^{-1} to the right now gives

$$\begin{aligned} \text{tr} \bar{E}E\varepsilon(k)\nu(k)^* &= \text{tr} \frac{1}{2\pi j} \oint \frac{1}{UU_*TDD_*} \tilde{\mathbf{V}}_* \tilde{\mathbf{V}} \\ &\quad \left\{ z^{-m} \mathbf{S} \tilde{\mathbf{C}} \tilde{\mathbf{C}}_* \hat{\mathbf{B}}_{o*} - T \mathbf{R} \mathbf{A}^{-1} \mathbf{N}^{-1} \beta \beta_* \mathbf{N}_*^{-1} \right\} \\ &\quad \times \mathbf{A}_*^{-1} \mathbf{M}_* \frac{dz}{z}. \end{aligned} \quad (\text{C.8})$$

To make (C.8) zero, all poles inside $|z| = 1$ are eliminated. This is achieved if, in every element of the integrand, all such poles are cancelled by zeros. We first cancel what can be cancelled by means of \mathbf{R} directly. Thus

$$\mathbf{R} = \frac{1}{T} \tilde{\mathbf{V}}^{-1} \mathbf{Q} \beta \beta^{-1} \mathbf{N} \mathbf{A} \quad (\text{C.9})$$

where $\mathbf{Q}(z^{-1})$ is undetermined. Inserting (C.9) into (C.8) gives $\text{tr} \bar{E}E\varepsilon(k)\nu(k)^*$ as

$$\begin{aligned} &\text{tr} \frac{1}{2\pi j} \oint \frac{1}{UU_*TDD_*} \tilde{\mathbf{V}}_* \\ &\quad \left\{ z^{-m} \tilde{\mathbf{V}} \mathbf{S} \tilde{\mathbf{C}} \tilde{\mathbf{C}}_* \hat{\mathbf{B}}_{o*} \mathbf{N}_* - \mathbf{Q} \beta_* \right\} \mathbf{N}_*^{-1} \mathbf{A}_*^{-1} \mathbf{M}_* \frac{dz}{z}. \end{aligned} \quad (\text{C.10})$$

Now, $U, T, D, \mathbf{A}, \mathbf{N}$ are all stable, so they have zeros only inside $|z| = 1$. The poles of \mathbf{M} are inside $|z| = 1$. This means that $U_*^{-1}, D_*^{-1}, \mathbf{A}_*^{-1}, \mathbf{N}_*^{-1}$, and \mathbf{M}_* have all their poles outside $|z| = 1$. No poles will thus exist inside $|z| = 1$ in (C.10), if and only if

$$z^{-m} \tilde{\mathbf{V}} \mathbf{S} \tilde{\mathbf{C}} \tilde{\mathbf{C}}_* \hat{\mathbf{B}}_{o*} \mathbf{N}_* - \mathbf{Q} \beta_* = z \mathbf{L}_* \mathbf{U} \mathbf{T} \mathbf{D} \mathbf{I}_p \quad (\text{C.11})$$

for some polynomial matrix $\mathbf{L}_*(z)$. This is (3.11), if q is substituted for z . The filter (C.9) coincides with (3.10). Necessity follows because choices of \mathbf{R} other than (C.9) correspond to $\nu(k) \neq 0$ in (C.3).

Unique solvability of (3.11) is demonstrated as follows. The Diophantine equation will always have one or several solutions, since the invariant polynomials of $\mathbf{U} \mathbf{T} \mathbf{D} \mathbf{I}_p$ are all

able, while those of β_* are all unstable. Thus, there exist no common invariant factors. Let (Q_0, L_{0*}) be one solution pair. Every solution to (3.11) can then be expressed as

$$(Q, L_*) = (Q_0 - XqUTDI_p, L_{0*} + X\beta_*)$$

where the polynomial matrix $X(q, q^{-1})$ is undetermined. Now, Q is required to be causal, so it can not have any positive powers of q as arguments, while L_* must contain no negative powers of q , to assure optimality. Thus, $X(q, q^{-1}) = 0$ is the only choice. We conclude that the solution to (3.11) is unique.

The degrees (3.12) are determined by the requirement that the maximum powers of q^{-1} and q are covered on both sides of (3.11). They assure that the number of unknowns equal the number of equations in the corresponding linear system of equations. For details, see [1] or [3].

The minimal average estimation error, $J_{min} = \text{tr} \bar{E}E(\varepsilon(k))_{min}$, is obtained as follows. First insert (3.10) into the criterion (2.3), use Parseval's formula, take expectation and use (3.1), (3.9) and (3.5) in this order. Then we obtain

$$\begin{aligned} J_{min} = & \text{tr} \frac{1}{2\pi j} \oint_{|z|=1} \frac{1}{UU^*TT^*DD_*} \tilde{V}^* \tilde{V} \left\{ S\tilde{C}\tilde{C}_*S_* \right. \\ & - S\tilde{C}\tilde{C}_* \hat{B}_{0*} A_*^{-1} R_* T_* z^{-m} - z^m T^* R A^{-1} \hat{B}_0 \tilde{C}\tilde{C}_* S_* \\ & \left. + T^* R A^{-1} N^{-1} \beta \beta_* N_*^{-1} A_*^{-1} R_* T_* \right\} \frac{dz}{z}. \quad (C.12) \end{aligned}$$

Now, the use of (C.9), $\text{tr} \tilde{V}^* \tilde{V} \{ \dots \} = \text{tr} \tilde{V} \{ \dots \} \tilde{V}^*$ and completing the square gives

$$\begin{aligned} J_{min} = & \text{tr} \frac{1}{2\pi j} \oint_{|z|=1} \frac{1}{UU^*TT^*DD_*} \tilde{V} S\tilde{C}\tilde{C}_* S_* \tilde{V}^* \\ & + \left(z^{-m} \tilde{V} S\tilde{C}\tilde{C}_* \hat{B}_{0*} N_* \beta_*^{-1} - Q \right) \\ & \times \left(\beta^{-1} N \hat{B}_0 \tilde{C}\tilde{C}_* S_* \tilde{V}^* z^m - Q_* \right) \\ & - \tilde{V} S\tilde{C}\tilde{C}_* \hat{B}_{0*} N_* \beta_*^{-1} \beta^{-1} N \hat{B}_0 \tilde{C}\tilde{C}_* S_* \tilde{V}^* \left\} \frac{dz}{z}. \end{aligned}$$

Finally, use Diophantine equation (C.11) in the middle term and rearrange the terms to obtain expression (3.13).

REFERENCES

- [1] A. Ahlén and M. Sternad, "Wiener filter design using polynomial equations," *IEEE Trans. Signal Processing*, vol. 39, pp. 2387–2399, 1991.
- [2] —, "Optimal filtering problems," in *Polynomial Methods in Optimal Control and Filtering*, K. Hunt, Ed. London: Peter Peregrinus, 1993.
- [3] —, "Derivation and design of Wiener filters using polynomial equations," *Control and Dynamic Systems*, vol. 64: *Stochastic Techniques in Digital Signal Processing Systems*, C. T. Leondes, Ed. New York: Academic, 1994, pp. 353–418.
- [4] P. Balaban and J. Salz, "Optimum diversity combining and equalization in digital data transmission with applications to cellular mobile radio—Part I: Theoretical considerations," *IEEE Trans. Communications*, vol. 40, pp. 885–894, 1992.
- [5] B. Bernhardsson and M. Sternad, "Feedforward control is dual to deconvolution," *Int. J. Contr.*, vol. 57, pp. 393–405, 1993.
- [6] B. Bernhardsson, "Robust stochastic performance optimization," IFAC World Congress, Sydney, Australia, July 1993, vol. 8, pp. 67–70. Also in *Topics in Digital and Robust Control of Linear Systems*. Ph.D. dissertation, TFRT-1039-SE, Dept. of Automatic Control, Lund Institute of Technology, Lund, Sweden, 1992.
- [7] P. Bolzern, P. Colaneri, and G. De Nicolao, "Optimal robust filtering for linear systems subject to time-varying parameter perturbations," in *Proc. 32nd Conf. Decis. Contr.*, San Antonio, TX, Dec. 1993, pp. 1018–1023.
- [8] B. Carlsson, A. Ahlén, and M. Sternad, "Optimal differentiation based on stochastic signal models," *IEEE Trans. Signal Processing*, vol. 39, pp. 341–353, 1991.
- [9] R. C. Chung and P. R. Bélanger, "Minimum-sensitivity filter for linear time-invariant stochastic systems with uncertain parameters," *IEEE Trans. Automat. Contr.*, vol. 21, pp. 98–100, 1976.
- [10] J. A. D'Appolito and C. E. Hutchinson, "A minimax approach to the design of low sensitivity state estimators," *Automatica*, vol. 8, pp. 599–608, 1972.
- [11] G. C. Goodwin, M. Gevers, and B. Ninness, "Quantifying the error in estimated transfer functions with application to model order selection," *IEEE Trans. Automat. Contr.*, vol. 27, pp. 913–928, 1992.
- [12] G. C. Goodwin and M. E. Salgado, "A stochastic embedding approach for quantifying uncertainty in the estimation of restricted complexity models," *Int. J. Adaptive Contr. Signal Processing*, vol. 3, pp. 333–356, 1989.
- [13] M. J. Grimble, "Wiener and Kalman filters for systems with random parameters," *IEEE Trans. Automat. Contr.*, vol. 29, pp. 552–554, 1984.
- [14] M. J. Grimble and A. ElSayed, "Solution to the \mathcal{H}_∞ optimal linear filtering problem for discrete-time systems," *IEEE Trans. Signal Processing*, vol. 38, pp. 1092–1104, 1990.
- [15] W. F. Haddad and D. S. Bernstein, "The optimal projection equations for reduced-order discrete-time state estimation for linear systems with multiplicative noise," *Syst. Contr. Lett.*, vol. 8, pp. 381–388, 1987.
- [16] —, "Robust, reduced-order, nonstrictly proper state estimation via the optimal projection equations with guaranteed cost bounds," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 591–595, 1988.
- [17] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic, 1970.
- [18] J. Ježek and V. Kučera, "Efficient algorithm for matrix spectral factorization," *Automatica*, vol. 21, pp. 663–669, 1985.
- [19] S. A. Kassam and T. L. Lim, "Robust Wiener filters," *J. Franklin Inst.*, pp. 171–185, 1977.
- [20] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: a survey," *Proc. IEEE*, vol. 73, pp. 433–481, 1985.
- [21] V. Kučera, *Discrete Linear Control. The Polynomial Equations Approach*. New York: Wiley, 1979.
- [22] —, "Factorization of rational spectral matrices: a survey," in *Proc. IEE Control'91*, Edinburgh, Mar. 1991, pp. 1074–1078.
- [23] W. Lee, *Mobile Cellular Telecommunications Systems*. New York: McGraw-Hill, 1989.
- [24] C. T. Leondes and J. O. Pearson, "A minimax filter for systems with large plant uncertainties," *IEEE Trans. Automat. Contr.*, vol. 17, pp. 266–268, 1972.
- [25] E. Lindskog, M. Sternad, and A. Ahlén, "Designing decision feedback equalizers to be robust with respect to channel time variations," in *Proc. Nordic Radio Society Symp. Disturbance Resistant Radio and Radar*, Uppsala, Sweden, Nov. 10–11 1993.
- [26] C. J. Martin and M. Mintz, "Robust filtering and prediction for linear systems with uncertain dynamics: a game-theoretic approach," *IEEE Trans. Automat. Contr.*, vol. 28, pp. 888–896, 1983.
- [27] R. H. Milocco, "Optimal design of feedforward regulators for multi-variable systems," in *Proc. Automat. Contr. Conf.*, Chicago, IL, June 24–26, 1992, pp. 780–784.
- [28] G. Moustakides and S. A. Kassam, "Robust Wiener filters for random signals in correlated noise," *IEEE Trans. Inform. Theory*, vol. 29, pp. 614–619, 1983.
- [29] K. Öhrn, "Design of multivariable cautious Wiener filters: A probabilistic approach," Licentiate thesis, Dept. Technology, Uppsala University, Uppsala, Sweden, in preparation.
- [30] I. R. Petersen and D. C. McFarlane, "Robust state estimation for uncertain systems," in *Proc. 30th Conf. Decis. Contr.*, Brighton, England, Dec. 1991, pp. 2630–2631.
- [31] H. V. Poor, "On robust Wiener filtering," *IEEE Trans. Automat. Contr.*, vol. 25, pp. 531–536, 1980.
- [32] J. L. Speyer and D. E. Gustafson, "An approximation method for estimation in linear systems with parameter uncertainty," *IEEE Trans. Automat. Contr.*, vol. 20, pp. 354–359, 1975.
- [33] M. Sternad and A. Ahlén, "The structure and design of realizable decision feedback equalizers for IIR channels with colored noise," *IEEE Trans. Inform. Theory*, vol. 36, pp. 848–858, 1990.
- [34] M. Sternad, A. Ahlén, and E. Lindskog, "Robust decision feedback equalization," in *Proc. ICASSP'93, Int. Conf. Acoustics, Speech and Signal Processing*, Minneapolis, MN, vol. III, April 26–30, 1993, pp. 555–558.

- [35] M. Sternad and A. Ahlén, "A novel derivation methodology for polynomial-LQ controller design" *IEEE Trans Automat Contr*, vol 38, pp 116-121, 1993
- [36] ———, "Robust filtering and feedforward control based on probabilistic descriptions of model errors," *Automatica*, vol 29, pp 661-679, 1993
- [37] M. Sternad, K. Öhrn and A. Ahlén, "Robust H_2 filtering for structured uncertainty: the performance of probabilistic and minimax schemes," Rep. UPTec 94090, Systems and Control Group, Uppsala University, submitted to the European Control Conference, Rome, Sept 1995
- [38] K. S. Vastola and H. V. Poor, "Robust Wiener-Kolmogorov theory," *IEEE Trans Inform Theory*, vol 30, pp 316-327, 1984
- [39] L. Xie and Y. C. Soh, "Robust Kalman filtering for uncertain systems," *Syst Contr Lett*, vol 22, pp 123-129, 1994
- [40] L. Xie, Y. C. Soh, and C. E. de Souza, "Robust Kalman filtering for uncertain discrete time systems" *IEEE Trans Automat Contr*, vol 39, pp 1310-1314, June 1994



Kenth Öhrn (S'92) was born in Västerås, Sweden, in 1960. He received the M.Sc. degree in engineering physics in 1991 from the Institute of Technology, Uppsala University, Sweden. He is working toward the Swedish Licentiate degree from the same institution.

He works on control problems at ABB Traction AB, Electrical Division in Västerås, Sweden. His current research interests include robust deconvolution and robust feedforward control.



Anders Ahlén (S'80-M'89-SM'90) received the Ph.D. degree in automatic control in 1986 from the Institute of Technology, Uppsala University, Sweden.

Dr. Ahlén held positions as Assistant and Associate Professor of Automatic Control at Uppsala University from 1984-1989 and 1989-1992, respectively. He spent 1991 with the Department of Electrical and Computer Engineering, University of Newcastle, Australia. From July 1992, he has been Associate Professor of Signal Processing at the Systems and Control Group, Uppsala University.

His research interests are currently focused on the robust and adaptive systems in the area of signal processing, communications, and control.



Mikael Sternad (S'83-M'86-SM'90) received the M.S. degree in engineering physics in 1981 and the Ph.D. degree in automatic control in 1987, both from the Institute of Technology, Uppsala University, Sweden.

Dr. Sternad is an Associate Professor at the University of Uppsala, Sweden. He has worked with the Systems and Control Group at Uppsala from 1981-1986 as teaching assistant and from 1987 as Assistant Professor and Associate Professor.

His current research interests include the use of polynomial equations approach on problems in control estimation, digital communications, and the design of adaptive tracking algorithms for systems with fast time variations.

A New Model for Control of Systems with Friction

C. Canudas de Wit, Associate, IEEE, H. Olsson, Student Member, IEEE, K. J. Åström, Fellow, IEEE, and P. Lischinsky

Abstract—In this paper we propose a new dynamic model for friction. The model captures most of the friction behavior that has been observed experimentally. This includes the Stribeck effect, hysteresis, spring-like characteristics for stiction, and varying break-away force. Properties of the model that are relevant to control design are investigated by analysis and simulation. New control strategies, including a friction observer, are explored, and stability results are presented.

I. INTRODUCTION

FRICITION is an important aspect of many control systems both for high quality servo mechanisms and simple pneumatic and hydraulic systems. Friction can lead to tracking errors, limit cycles, and undesired stick-slip motion. Control strategies that attempt to compensate for the effects of friction, without resorting to high gain control loops, inherently require a suitable friction model to predict and to compensate for the friction. These types of schemes are therefore named model-based friction compensation techniques. A good friction model is also necessary to analyze stability, predict limit cycles, find controller gains, perform simulations, etc. Most of the existing model-based friction compensation schemes use classical friction models, such as Coulomb and viscous friction. In applications with high precision positioning and with low velocity tracking, the results are not always satisfactory. A better description of the friction phenomena for low velocities and especially when crossing zero velocity is necessary. Friction is a natural phenomenon that is quite hard to model, and it is not yet completely understood. The classical friction models used are described by static maps between velocity and friction force. Typical examples are different combinations of Coulomb friction, viscous friction, and Stribeck effect [1]. The latter is recognized to produce a destabilizing effect at very low velocities. The classical models explain neither hysteretic behavior when studying friction for nonstationary velocities nor variations in the break-away force with the experimental condition nor small displacements that occur at the contact interface during stiction. The latter

very much resembles that of a connection with a stiff spring with damper and is sometimes referred to as the Dahl effect. Later studies (see, e.g., [1], [2]) have shown that a friction model involving dynamics is necessary to describe the friction phenomena accurately.

A dynamic model describing the spring-like behavior during stiction was proposed by Dahl [3]. The Dahl model is essentially Coulomb friction with a lag in the change of friction force when the direction of motion is changed. The model has many nice features, and it is also well understood theoretically. Questions such as existence and uniqueness of solutions and hysteresis effects were studied in an interesting paper by Bliman [4]. The Dahl model does not, however, include the Stribeck effect. An attempt to incorporate this into the Dahl model was done in [5] where the authors introduced a second-order Dahl model using linear space invariant descriptions. The Stribeck effect in this model is only transient, however, after a velocity reversal and is not present in the steady-state friction characteristics. The Dahl model has been used for adaptive friction compensation [6], [7], with improved performance as the result. There are also other models for dynamic friction. Armstrong-Hélouvy proposed a seven parameter model in [1]. This model does not combine the different friction phenomena but is in fact one model for stiction and another for sliding friction. Another dynamic model suggested by Rice and Ruina [8] has been used in connection with control by Dupont [9]. This model is not defined at zero velocity. In this paper we will propose a new dynamic friction model that combines the stiction behavior, i.e., the Dahl effect, with arbitrary steady-state friction characteristics which can include the Stribeck effect. We also show that this model is useful for various control tasks.

II. A NEW FRICTION MODEL

The qualitative mechanisms of friction are fairly well understood (see, e.g., [1]). Surfaces are very irregular at the microscopic level and two surfaces therefore make contact at a number of asperities. We visualize this as two rigid bodies that make contact through elastic bristles. When a tangential force is applied, the bristles will deflect like springs which gives rise to the friction force; see Fig. 1.

If the force is sufficiently large some of the bristles deflect so much that they will slip. The phenomenon is highly random due to the irregular forms of the surfaces. Haessig and Friedland [10] proposed a bristle model where the random behavior was captured and a simpler reset-integrator model which describes the aggregated behavior of the bristles. The model we propose is also based on the average behavior of

Manuscript received August 27, 1993, revised February 15, 1994. Recommended by Associate Editor, T. A. Posbergh. This work was supported in part by the Swedish Research Council for Engineering Sciences (TFR) Contract 91-721, the French National Scientific Research Council (CNRS), and the EU Human Capital and Mobility Network on Nonlinear and Adaptive Control ERBCHRXCT 93-0380.

C. Canudas de Wit and P. Lischinsky are with Laboratoire d'Automatique de Grenoble, URA CNRS 228, ENSIEG-INPG, B.P. 46, 38402, Grenoble, France.

H. Olsson and K. J. Åström are with the Department of Automatic Control, Lund Institute of Technology, Box 118, S-221 00 Lund, Sweden.

IEEE Log Number 9408272.

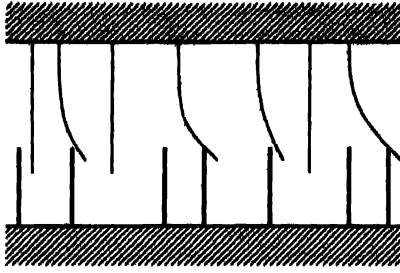


Fig. 1. The friction interface between two surfaces is thought of as a contact between bristles. For simplicity the bristles on the lower part are shown as being rigid.

the bristles. The average deflection of the bristles is denoted by z and is modeled by

$$\frac{dz}{dt} = v - \frac{|v|}{g(v)} z \quad (1)$$

where v is the relative velocity between the two surfaces. The first term gives a deflection that is proportional to the integral of the relative velocity. The second term asserts that the deflection z approaches the value

$$z_{ss} = \frac{v}{|v|} g(v) = g(v) \operatorname{sgn}(v) \quad (2)$$

in steady state, i.e., when v is constant. The function g is positive and depends on many factors such as material properties, lubrication, temperature. It need not be symmetrical. Direction dependent behavior can therefore be captured. For typical bearing friction, $g(v)$ will decrease monotonically from $g(0)$ when v increases. This corresponds to the Stribeck effect. The friction force generated from the bending of the bristles is described as

$$F = \sigma_0 z + \sigma_1 \frac{dz}{dt}$$

where σ_0 is the stiffness and σ_1 a damping coefficient. A term proportional to the relative velocity could be added to the friction force to account for viscous friction so that

$$F = \sigma_0 z + \sigma_1 \frac{dz}{dt} + \sigma_2 v. \quad (3)$$

The model given by (1) and (3) is characterized by the function g and the parameters σ_0 , σ_1 and σ_2 . The function $\sigma_0 g(v) + \sigma_2 v$ can be determined by measuring the steady-state friction force when the velocity is held constant. A parameterization of g that has been proposed to describe the Stribeck effect is

$$\sigma_0 g(v) = F_C + (F_S - F_C) e^{-(v/v_s)^2} \quad (4)$$

where F_C is the Coulomb friction level, F_S is the level of the stiction force, and v_s is the Stribeck velocity; see [1]. With this description the model is characterized by six parameters σ_0 , σ_1 , σ_2 , F_C , F_S , and v_s . It follows from (2)–(4) that for steady-state motion the relation between velocity and friction force is given by

$$\begin{aligned} F_{ss}(v) &= \sigma_0 g(v) \operatorname{sgn}(v) + \sigma_2 v \\ &= F_C \operatorname{sgn}(v) + (F_S - F_C) e^{-(v/v_s)^2} \operatorname{sgn}(v) + \sigma_2 v. \end{aligned}$$

Note, however, that when velocity is not constant, the dynamics of the model will be very important and give rise to different types of phenomena. This will be discussed in Section IV.

Relation to the Dahl Model

The model reduces to the Dahl model if $g(v) = F_C/\sigma_0$, and $\sigma_1 = \sigma_2 = 0$. Equations (1) and (3) then give

$$\frac{dF}{dt} = \sigma_0 \frac{dz}{dt} = \sigma_0 v \left(1 - \frac{F}{F_C} \operatorname{sgn}(v) \right). \quad (5)$$

Dahl actually suggested the more general model

$$\frac{dF}{dt} = \sigma_0 v |1 - \frac{F}{F_C} \operatorname{sgn}(v)|^n \operatorname{sgn} \left(1 - \frac{F}{F_C} \operatorname{sgn}(v) \right)$$

see [11]. Most references to Dahl's work, however, do use the simpler model (5). Dahl's model accounts for Coulomb friction but it does not describe the Stribeck effect.

An Extension of the Dahl Model

An attempt to extend Dahl's model to include the Stribeck effect was made by Bliman and Sorine [5]. They replaced the time variable t by a space variable s through the transformation

$$s = \int_0^t |v(\tau)| d\tau.$$

Equation (5) then becomes

$$\frac{dF}{ds} = -\sigma_0 \frac{F}{F_C} + \sigma_0 \operatorname{sgn}(v) \quad (6)$$

which is a linear first-order system if $\operatorname{sgn}(v)$ is regarded as an input. Bliman and Sorine then replaced (6) by the second-order model

$$\frac{d^2 F}{ds^2} + 2\zeta\omega \frac{dF}{ds} + \omega^2 F = \omega^2 F_C \operatorname{sgn}(v)$$

to imitate the Stribeck effect with an overshoot in the response to sign changes in the velocity. This model, however, will only give a spatially transient Stribeck effect after a change of the direction of motion. The Stribeck effect is not present in the steady-state relation between velocity and friction force.

III. MODEL PROPERTIES

The properties of the model given by (1) and (3) will now be explored. To capture the intuitive properties of the bristle model in Fig. 1, the deflection z should be finite. This is indeed the case because we have the following property.

Property 1: Assume that $0 < g(v) \leq a$. If $|z(0)| \leq a$ then $|z(t)| \leq a \forall t \geq 0$.

Proof: Let $V = z^2/2$, then the time derivative of V evaluated along the solution of (1) is

$$\begin{aligned} \frac{dV}{dt} &= z \left(v - \frac{|v|}{g(v)} z \right) \\ &= -|v| |z| \left(\frac{|z|}{g(v)} - \operatorname{sgn}(v) \operatorname{sgn}(z) \right) \end{aligned}$$

TABLE I
PARAMETER VALUES USED IN ALL SIMULATIONS

Parameter	Value	Unit
σ_0	10^5	[N/m]
σ_1	$\sqrt{10^5}$	[Ns/m]
σ_2	0.4	[Ns/m]
F_C	1	[N]
F_S	1.5	[N]
v_S	0.001	[m/s]

The derivative $\frac{dV}{dt}$ is negative when $|z| > g(v)$. Since $g(v)$ is strictly positive and bounded by a , we see that the set $\Omega = \{z : |z| \leq a\}$ is an invariant set for the solutions of (1), i.e., all the solutions of $z(t)$ starting in Ω remain there.

Dissipativity

Intuitively we may expect that friction will dissipate energy. Since our model given by (1) and (3) is dynamic, there may be phases where friction stores energy and others where it gives energy back. It can be proven that the map $\varphi : v \mapsto z$ is dissipative for our model. For more details on the concepts and definitions concerning dissipative systems, see [12].

Property 2: The map $\varphi : v \mapsto z$, as defined by (1), is dissipative with respect to the function $V(t) = \frac{1}{2}z^2(t)$, i.e.,

$$\int_0^t z(\tau)v(\tau) d\tau \geq V(t) - V(0).$$

Proof: It follows from (1) that

$$\begin{aligned} zv &= z \frac{dz}{dt} + \frac{|v|}{g(v)} z^2 \\ &\geq z \frac{dz}{dt}. \end{aligned}$$

Hence

$$\int_0^t z(\tau)v(\tau) d\tau \geq \int_0^t z(\tau) \frac{dz(\tau)}{d\tau} d\tau \geq V(t) - V(0).$$

Linearization in Stiction Regime

To get some insight into the behavior of the model in the stiction regime we will consider a mass m in contact with a fixed horizontal surface. Let x be the coordinate of the mass, i.e., $v = dx/dt$. The equation of motion becomes

$$m \frac{d^2x}{dt^2} = -F = -\sigma_0 z - \sigma_1 \frac{dz}{dt} - \sigma_2 \frac{dx}{dt} \quad (7)$$

where z is given by (1). Linearizing (1) around $z = 0$ and $v = 0$ we get

$$\frac{dz}{dt} = \frac{dx}{dt} \quad (8)$$

Inserting (8) into (7) gives

$$m \frac{d^2x}{dt^2} + (\sigma_1 + \sigma_2) \frac{dx}{dt} + \sigma_0 x = 0. \quad (9)$$

This shows that the system behaves like a damped second-order system. Notice that the bristle stiffness, σ_0 , is usually

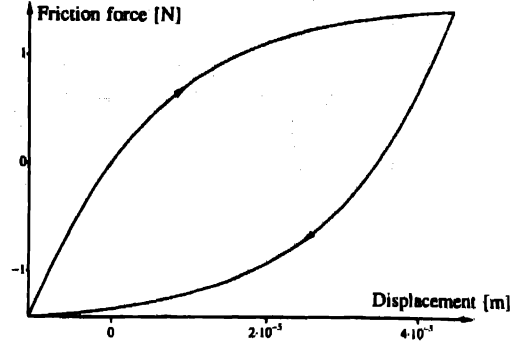


Fig. 2. Presliding displacement as described by the model. The simulation was started with zero initial conditions.

very large, and therefore it is essential to have $\sigma_1 \neq 0$ to have a sufficiently damped motion. The viscous friction coefficient, σ_2 , is normally not sufficiently large to provide good damping.

IV. DYNAMICAL MODEL BEHAVIOR

As a preliminary assessment of the model we will investigate its behavior in some typical cases. They correspond to standard experiments that have been performed. In all the simulations the function g has been parameterized according to (4) and the parameter values in Table I have been used. The parameter values have to some extent been based on experimental results [1]. The stiffness σ_0 was chosen to give a presliding displacement of the same magnitude as reported in various experiments. The value of the damping coefficient σ_1 was chosen to give a damping of $\zeta = 0.5$ for the linearized equation (9) with a unit mass. The Coulomb friction level F_C corresponds to a friction coefficient $\mu \approx 0.1$ for a unit mass, and F_S gives a 50% higher friction for very low velocities. The viscous friction σ_2 and the Stribeck velocity v_s are also of the same order of magnitude as given in [1].

The different behaviors shown in the following subsections cannot be attributed to single parameters but rather to the behavior of the nonlinear differential equation (1) and the shape of the function g . The presliding displacement and the varying break-away force are due to the dynamics. This behavior is also present in the Dahl model. The Stribeck shape of g together with the dynamics give rise to the type of hysteresis observed in the subsection on frictional lag.

A. Presliding Displacement

Courtney-Pratt and Eisner have shown that friction behaves like a spring if the applied force is less than the break-away force. If a force is applied to two surfaces in contact there will be a displacement. A simulation was performed to investigate if our model captures this phenomenon. An external force was applied to a unit mass subjected to friction. The applied force was slowly ramped up to 1.425 N which is 95% of F_S . The force was then kept constant for a while and later ramped down to the value -1.425 N, where it was kept constant and then ramped up to 1.425 N again. The results of the simulation are shown in Fig. 2 where the friction force is shown as a function of displacement. The behavior shown in Fig. 2 agrees qualitatively with the experimental results in [13].

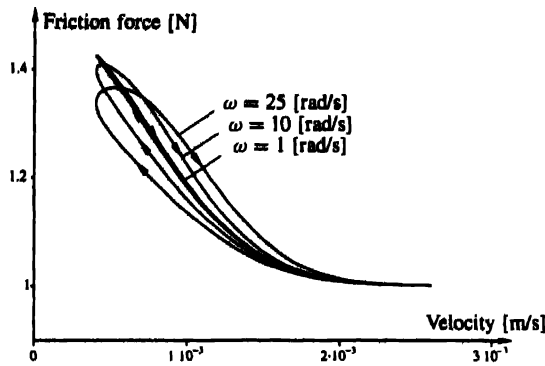


Fig. 3. Hysteresis in friction force with varying velocity. The velocity variation with the highest frequency shows the widest hysteresis loop

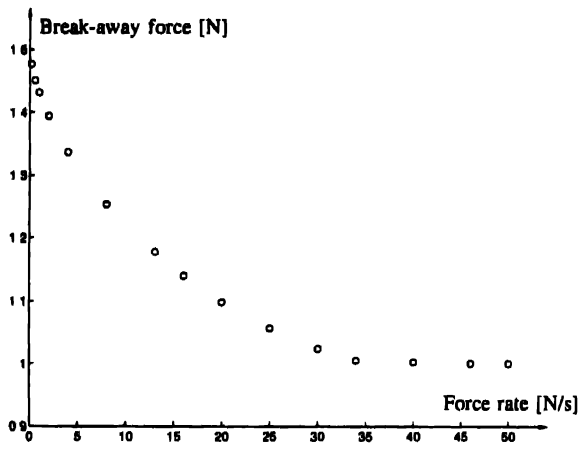


Fig. 4. Relation between break-away force and rate of increase of the applied force.

B. Frictional Lag

Hess and Soom [14] studied the dynamic behavior of friction when velocity is varied during unidirectional motion. They showed that there is hysteresis in the relation between friction and velocity. The friction force is lower for decreasing velocities than for increasing velocities. The hysteresis loop becomes wider at higher rates of the velocity changes. Hess and Soom explained their experimental results by a pure time delay in the relation between velocity and friction force. Fig. 3 shows a simulation of the Hess–Soom experiment using our friction model. The input to the friction model was the velocity which was changed sinusoidally around an equilibrium. The resulting friction force is given as a function of velocity in Fig. 3. Our model clearly exhibits hysteresis. The width of the hysteresis loop also increases with frequency. Our model thus captures the hysteretic behavior of real friction described in [14].

C. Varying Break-Away Force

The break-away force can be investigated through experiments with stick-slip motion. In [15] it is pointed out that in such experiments the dwell-time when sticking and the rate of increase of the applied force are always related and hence the effects of these factors cannot be separated. The experiment was therefore redesigned so that the time in stiction and the rate of increase of the applied force could be varied independently. The results showed that the break-away force

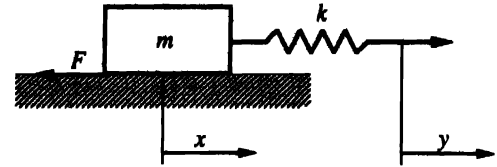


Fig. 5. Experimental setup for stick-slip motion.

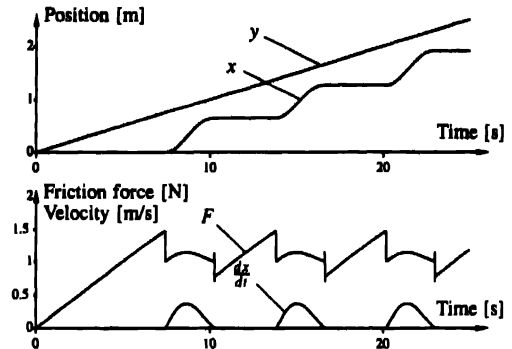


Fig. 6. Simulation of stick-slip motion

did depend on the rate of increase of the force but not on the dwell-time; see also [16]. Simulations were performed using our model to determine the break-away force for different rates of force application. Since the model is dynamic, a varying break-away force can be expected. A force applied to a unit mass was ramped up at different rates, and the friction force when the mass started to slide was determined. Note that since the model behavior in stiction is essentially that of a spring, there will be microscopic motion, i.e., velocity different from zero, as soon as a force is applied. The break-away force was therefore determined at the time where a sharp increase in the velocity could be observed. Fig. 4 shows the force at break-away as a function of the rate of increase of the applied force. The results agree qualitatively with the experimental results in [15] and [16].

D. Stick-Slip Motion

Stick-slip motion is a typical behavior for systems with friction. It is caused by the fact that friction is larger at rest than during motion. A typical experiment that may give stick-slip motion is shown in Fig. 5. A unit mass is attached to a spring with stiffness $k = 2$ N/m. The end of the spring is pulled with constant velocity, i.e., $dy/dt = 0.1$ m/s. Fig. 6 shows results of a simulation of the system based on the friction model in Section II. The mass is originally at rest and the force from the spring increases linearly. The friction force counteracts the spring force, and there is a small displacement. When the applied force reaches the break-away force, in this case approximately $\sigma_0 g(0)$, the mass starts to slide and the friction decreases rapidly due to the Stribeck effect. The spring contracts, and the spring force decreases. The mass slows down and the friction force increases because of the Stribeck effect and the motion stops. The phenomenon then repeats itself. In Fig. 6 we show the positions of the mass and the spring, the friction force and the velocity. Notice the highly irregular behavior of the friction force around the region where the mass stops.

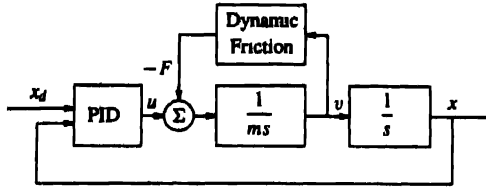


Fig. 7 Block diagram for the servo problem with PID controller

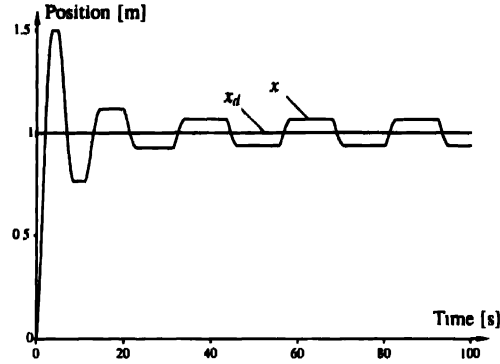


Fig. 8 Simulation of the PID position control problem in Fig. 7

V FEEDBACK CONTROL

To further illustrate the properties of our friction model we will investigate its application to some typical servo problems. First we will use it to show that it predicts limit cycle oscillations in servos with PID control. We will then use it to design observer based friction compensators.

A Limit Cycles Caused by Friction

It has been observed experimentally that friction may give rise to limit cycles in servo drives where the controller has integral action, for references see [2]. This phenomenon is often referred to as hunting.

Consider the linear motion of a mass m at position x . The equation of motion is

$$m \frac{d^2 x}{dt^2} = u - F \quad (10)$$

where $dx/dt = v$ is the velocity, F the friction force given by (3) and u the control force which is given by the PID controller

$$u = -K_v v - K_p (x - x_d) - K_i \int (x - x_d) \quad (11)$$

A block diagram of the system is shown in Fig. 7. In Fig. 8 we show the results of a simulation of the system. The friction parameters are given by Table 1, $m = 1$ and the controller parameters are $K_v = 6$, $K_p = 3$, and $K_i = 4$. The reference position is chosen as $x_d = 1$. The model clearly predicts limit cycles as have been observed experimentally in systems of this type.

B Friction Compensation

Only linear feedback from the position was used in the PID control law (11). Knowledge about friction was not used. It is of course more appealing to make a model-based control that uses the model to predict the friction to compensate for

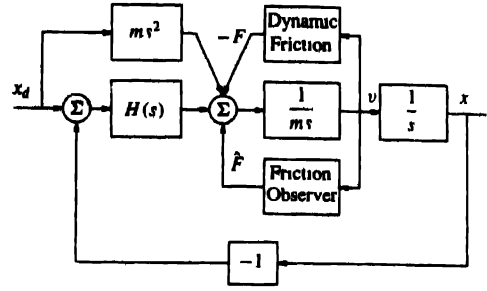


Fig. 9 Block diagram for the position control problem using a friction observer

it. Since the model is dynamic and has an unmeasurable state, some kind of observer is necessary. This will be discussed next.

Position Control with a Friction Observer Consider the problem of position tracking for the process (10). Assume that the parameters σ_0 , σ_1 , and σ_2 , and the function q in the friction model are known. The state z is, however, not measurable and hence has to be observed to estimate the friction force. For this we use a nonlinear friction observer given by

$$\frac{dz}{dt} = v - \frac{|v|}{q(v)} z - k \epsilon, \quad k > 0 \quad (12)$$

$$\hat{F} = \sigma_0 z + \sigma_1 \frac{dz}{dt} + \sigma_2 v \quad (13)$$

and the following control law

$$u = -H(v) \epsilon + \hat{F} + m \frac{d^2 x_d}{dt^2} \quad (14)$$

where $\epsilon = x - x_d$ is the position error and x_d is the desired reference which is assumed to be twice differentiable. The term $k \epsilon$ in the observer is a correction term from the position error. The closed-loop system is represented by the block diagram in Fig. 9. With the observer based friction compensation, we achieve position tracking as shown in the following theorem.

Theorem 1 Consider system (10) together with the friction model (1) and (3), friction observer (12) and (13), and control law (14). If $H(v)$ is chosen such that

$$G(s) = \frac{\sigma_1 s + \sigma_0}{m s^2 + H(s)}$$

is strictly positive real (SPR) then the observer error, $F - \hat{F}$, and the position error, ϵ , will asymptotically go to zero.

Proof The control law yields the following equations

$$\epsilon = \frac{1}{m s^2 + H(s)} (-\hat{F}) = \frac{\sigma_1 s + \sigma_0}{m s^2 + H(s)} (-z) = -G(s) z$$

$$\frac{dz}{dt} = -\frac{|v|}{q(v)} \tilde{z} + k \epsilon$$

where $\tilde{F} = F - \hat{F}$ and $z = z - \tilde{z}$. Now introduce

$$V = \xi^T P \xi + \frac{z^2}{k}$$

as a Lyapunov function and

$$\begin{aligned} \frac{d\xi}{dt} &= A\xi + B(-z) \\ \dot{\epsilon} &= C\xi \end{aligned}$$

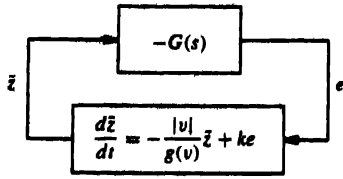


Fig. 10. The block diagram in Fig. 9 redrawn with e and \tilde{z} as outputs of a linear and a nonlinear block, respectively.

which is a state-space representation of $G(s)$. Since $G(s)$ is SPR [17] it follows from the Kalman–Yakubovitch Lemma [17] that there exist matrices $P = P^T > 0$ and $Q = Q^T > 0$ such that

$$\begin{aligned} A^T P + P A &= -Q \\ P B &= C^T. \end{aligned}$$

Now

$$\begin{aligned} \frac{dV}{dt} &= -\xi^T Q \xi - 2\xi^T P B \tilde{z} + \frac{2}{k} \tilde{z} \frac{d\tilde{z}}{dt} \\ &= -\xi^T Q \xi - 2c\tilde{z} + \frac{2}{k} \tilde{z} \left(-\frac{|v|}{g(v)} \tilde{z} + k e \right) \\ &= -\xi^T Q \xi - \frac{2}{k} \frac{|v|}{g(v)} \tilde{z}^2 \\ &\leq -\xi^T Q \xi. \end{aligned}$$

The radial unboundedness of V together with the semi-definiteness of dV/dt implies that the states are bounded. We can now apply LaSalle's theorem to see that $\xi \rightarrow 0$ and $\tilde{z} \rightarrow 0$ which means that both e and \tilde{F} tends to zero and the theorem is proven.

The theorem can also be understood from the following observations. By introducing the observer we get a dissipative map from e to \tilde{z} and by adding the friction estimate to the control signal, the position error will be the output of a linear system operating on \tilde{z} . This means that we have an interconnection of a dissipative system and a linear SPR system as seen in Fig. 10. Such a system is known to be asymptotically stable.

Velocity Control: The same type of observer-based control can be used for velocity control. For this control problem the controller is changed to

$$\begin{aligned} u &= -H(s)e + \hat{F} + m \frac{dv_d}{dt} \\ \frac{d\tilde{z}}{dt} &= v - \frac{|v|}{g(v)} \tilde{z} - k(v - v_d), \quad k > 0 \\ \hat{F} &= \sigma_0 \tilde{z} + \sigma_1 \frac{d\tilde{z}}{dt} + \sigma_2 v \end{aligned} \quad (15)$$

where $v - v_d$ is the velocity error and v_d the desired velocity which is assumed to be differentiable. Velocity tracking is achieved as shown in the following theorem.

Theorem 2: Consider system (10) together with friction model (1) and (3) and observer based control law (15). If $H(s)$ is chosen such that

$$G(s) = \frac{\sigma_1 s + \sigma_0}{m s + H(s)}$$

is strictly positive real, then the observer error, $F - \hat{F}$, and the velocity error will asymptotically go to zero.

Proof: The theorem is proven in the same way as Theorem 1 after observing that the control law yields the following error equations

$$\begin{aligned} v - v_d &= \frac{1}{m s + H(s)} (-\hat{F}) = \frac{\sigma_1 s + \sigma_0}{m s + H(s)} (-\tilde{z}) = -G(s) \tilde{z} \\ \frac{d\tilde{z}}{dt} &= -\frac{|v|}{g(v)} \tilde{z} + k(v - v_d). \end{aligned}$$

This is again an interconnection of a dissipative system with \tilde{z} as its output and a linear SPR system with $v - v_d$ as its output.

To assume that the friction model and its parameters are known exactly is of course a strong assumption. Investigation of the sensitivity of the results to these assumptions is an interesting problem that is outside the scope of this paper. The accuracy required in the velocity measurement is a similar problem.

VI. CONCLUSIONS

A new dynamic model for friction has been presented. The model is simple yet captures most friction phenomena that are of interest for feedback control. The low velocity friction characteristics are particularly important for high performance pointing and tracking. The model can describe arbitrary steady-state friction characteristics. It supports hysteretic behavior due to frictional lag, spring-like behavior in stiction and gives a varying break-away force depending on the rate of change of the applied force. All these phenomena are unified into a first-order nonlinear differential equation. The model can readily be used in simulations of systems with friction.

Some relevant properties of the model have been investigated. The model was used to simulate position control of a servo with a PID controller. The simulations predict hunting as has been observed in applications of position control with integral action. The model has also been used to construct a friction observer and to perform friction compensation for position and velocity tracking. When the parameters are known the observer error and the control error will asymptotically go to zero. Sensitivity studies, parameter estimation and adaptation are natural extensions of this work.

REFERENCES

- [1] B. Armstrong-Hélouvy, *Control of Machines with Friction*. Boston, MA: Kluwer, 1991.
- [2] B. Armstrong-Hélouvy, P. Dupont, and C. Canudas de Wit, "A survey of models, analysis tools and compensation methods for the control of machines with friction," *Automatica*, vol. 30, no. 7, pp. 1083–1138, 1994.
- [3] P. Dahl, "A solid friction model," Aerospace Corp., El Segundo, CA, Tech. Rep. TOR-0158(3107-18)-1, 1968.
- [4] P.-A. Bliman, "Mathematical study of the Dahl's friction model," *European J. Mechanics. A/Solids*, vol. 11, no. 6, pp. 835–848, 1992.
- [5] P.-A. Bliman and M. Sorine, "Friction modelling by hysteresis operators Application to Dahl, Sticktion, and Strbeck effects," in *Proc Conf Models of Hysteresis*, Trento, Italy, 1991.
- [6] C. Walrath, "Adaptive bearing friction compensation based on recent knowledge of dynamic friction," *Automatica*, vol. 20, no. 6, pp. 717–727, 1984.
- [7] N. Ehrlich Leonard and P. Krishnaprasad, "Adaptive friction compensation for bi-directional low-velocity position tracking," in *Proc 31st Conf Decis Contr*, 1992, pp. 267–273.

- [8] J. R. Rice and A.L. Ruina, "Stability of steady frictional slipping," *J. Applied Mechanics*, vol. 50, no. 2, 1983.
- [9] P. Dupont, "Avoiding stick-slip through PD control," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 1094-1097, 1994.
- [10] D. Haessig and B. Friedland, "On the modeling and simulation of friction," in *Proc. 1990 Amer. Contr. Conf.*, San Diego, CA, 1990, pp. 1256-1261.
- [11] P. Dahl, "Solid friction damping of spacecraft oscillations," in *Proc. AIAA Guidance and Contr. Conf.*, Boston, MA, Paper 75-1104, 1975.
- [12] J. Willems, "Dissipative dynamical systems part I: General theory," *Arch. Rational Mech. Anal.*, vol. 45, pp. 321-51, 1972.
- [13] J. Courtney-Pratt and E. Eisner, "The effect of a tangential force on the contact of metallic bodies," in *Proc. Royal Society*, vol. A238, 1957, pp. 529-550.
- [14] D. P. Hess and A. Soom, "Friction at a lubricated line contact operating at oscillating sliding velocities," *J. Tribology*, vol. 112, pp. 147-152, 1990.
- [15] V. I. Johannes, M. A. Green, and C. A. Brockley, "The role of the rate of application of the tangential force in determining the static friction coefficient," *Wear*, vol. 24, no. 381-385, 1973.
- [16] R. S. H. Richardson and H. Nolle, "Surface friction under time-dependent loads," *Wear*, vol. 37, no. 1, pp. 87-101, 1976.
- [17] H. K. Khalil, *Nonlinear Systems*. New York: Macmillan, 1992.



Carlos Canudas de Wit (A-93) received the B.Sc. degree in electronics and communications from the Technologic of Monterrey, Mexico in 1980. He received the M.Sc. and the Ph.D. degrees in automatic control from the Polytechnic of Grenoble, France, in 1984 and 1987, respectively.

From 1981 to 1982 he worked as a Research Engineer at the Department of Electrical Engineering at the CINVESTAV-IPN in Mexico City. He was a Visiting Researcher in 1985 at Lund Institute of Technology, Sweden. Since 1987 he has been an

Associate Professor in the Department of Automatic Control, Polytechnic of Grenoble, where he teaches and conducts research in the area of adaptive and robot control. He wrote *Adaptive Control of Partially Known Systems: Theory and Applications* (Elsevier, 1988). He is an editor of *Advanced Robot Control* (Springer-Verlag) and also an associate editor of *IEEE TRANSACTIONS ON AUTOMATIC CONTROL*.



Henrik Olsson (S'91) received the M.Sc. degree in electrical engineering from Lund Institute of Technology, Lund, Sweden in 1989.

He spent the academic year 1989-1990 at the Department of Electrical Engineering at University of California, Santa Barbara. Since 1990 he has been with the Department of Automatic Control at Lund Institute of Technology where he is currently completing the Ph.D. degree. His main research interest is in control of nonlinear servosystems.



Karl Johan Åström (M'71-SM'77-F'79) received the Ph.D. degree in automatic control and mathematics from the Royal Institute of Technology (KTH), Stockholm, in 1960.

He has been Professor of Automatic Control at Lund Institute of Technology/Lund University since 1965. His research interests include broad aspects of automatic control, stochastic control, system identification, adaptive control, computer control, and computer-aided control engineering.

Dr. Åström has published five books and many papers. He is a member of the Royal Swedish Academy of Sciences, and the Royal Swedish Academy of Engineering Sciences (IVA). He has received many awards among them the Quazza medal from IFAC in 1987 and the IEEE Medal of Honor in 1993.



Pablo Lischinsky was born in Montevideo, Uruguay, on September 24, 1960. He received the B.S. degree and the M.S. degree in control engineering from the Escuela de Ingeniería de Sistemas, Universidad de Los Andes, Mérida, Venezuela, in 1985 and 1990, respectively. He received the M.S. degree in automatic control in 1993 from the Institut Nationale Polytechnique (INPG-ENSIEG), Laboratoire d'Automatique, Grenoble, France. Since 1990, he has been with the Department of Automatic Control at the Universidad

de Los Andes. Currently he is on leave, working on his Ph.D. dissertation at the Institut Nationale Polytechnique (INPG-ENSIEG). His current research interests are in adaptive control, identification, and computer control of mechanical systems.

Adaptive Nonlinear Design with Controller-Identifier Separation and Swapping

Miroslav Krstić, *Student Member, IEEE*, and Petar V. Kokotović, *Fellow, IEEE*

Abstract—We present a new adaptive nonlinear control design which achieves a complete controller-identifier separation. This modularity is made possible by a strong input-to-state stability property of the new controller with respect to the parameter estimation error and its derivative as inputs. These inputs are independently guaranteed to be bounded by the identifier. The new design is more flexible than the Lyapunov-based design because the identifier can employ any standard update law gradient and least-squares, normalized and unnormalized. A key ingredient in the identifier design and convergence analysis is a nonlinear extension of the well-known linear swapping lemma.

I. INTRODUCTION

THE estimation-based approach to adaptive control has been extremely successful in linear systems. In contrast to the Lyapunov-based approach, which restricts the choice of parameter update laws and controller structures, the estimation-based designs are versatile. For linear systems, any common update law and any stabilizing controller can be employed as long as the boundedness properties of the identifier are sufficient to allow a “certainty-equivalence” design of the controller. This versatility is of conceptual and practical importance. It is due to a modularity feature: the identifier module achieves its boundedness properties independently of the controller module.

Thanks to its versatility, the estimation-based approach unifies many diverse adaptive schemes. For linear systems, this unification, initiated by Egardt [5], was extended by Goodwin and Mayne [6].

Attempts to apply estimation-based designs to nonlinear systems have had only limited success. The nonlinearities were either matched [24], [2], [3] or severely restricted [27], [35], [9], [10], [39]. Otherwise the results were local, i.e., valid in regions which were not a priori verifiable. A cause for this difficulty is a fundamental difference between the instability phenomena in linear and nonlinear systems. The states of an unstable linear system remain bounded over any finite interval, so that there is enough time for the identifier to “catch up.” The situation is fundamentally different in a system with nonlinearities whose growth is faster than linear (x^2 , x_1x_2 , e^{x^2} , etc.). Even a small parameter estimation error may drive the state of such a nonlinear system to infinity in

finite time. This explains why estimation-based designs have been mostly for systems with linearly bounded nonlinearities. Typically, linear growth constraints had to be imposed not only on the plant nonlinearities, but also on those derived during the design.

The only nonlinear estimation-based results which go beyond the linear growth constraints were obtained by Praly *et al.* [29]–[33]. In [32] a unified framework of control Lyapunov functions was used to characterize relationships between nonlinear growth constraints and controller stabilizing properties. In the absence of matching conditions, all the nonlinear estimation-based schemes presented in [32] involved some growth restrictions.

In contrast to the difficulties experienced by the estimation-based designs, the new recursive Lyapunov-based designs for systems in the parametric-strict-feedback form [12], [8], [19], [36] and the output-feedback form [21], [22], [14] were successful in achieving global boundedness and tracking without any restrictions on nonlinearities. However, these designs do not allow any flexibility in the choice of the parameter update law, excluding, for example, the least-squares update laws.

In spite of the previous difficulties with nonlinear estimation-based approaches, their flexibility and modularity motivate us to pursue their development. Since the independence of the identifier is not sufficient for modularity, we place the burden of the task of boundedness on the controller. For parametric-strict-feedback systems we seek (and find!) nonlinear controllers which guarantee boundedness in the presence of bounded parameter uncertainty. More precisely, we consider the parameter estimation error and its derivative as two independent disturbance inputs and design controllers which achieve input-to-state stability [37] (ISS) with respect to those inputs. In addition to such ISS-controllers, we also design weaker SG-controllers which only provide a small gain property and are presented for comparison with linear designs.

These new controllers create a possibility for a complete identifier-controller modularity. The remaining task is to design identifiers with guaranteed boundedness properties. A key ingredient in the identifier design and convergence analysis in this paper is our nonlinear extension of the well known linear swapping lemma [25]. Various forms of swapping were also used in most of the early nonlinear estimation-based results [27], [24], [29], [30], [35], [2], [9], [10], [39]. The identifiers in this paper are based on two different parametric models: the plant model and the error system. They allow a wide variety of update laws—gradient and least-squares, normalized and unnormalized.

Manuscript received March 26, 1993; revised December 7, 1993 and May 31, 1994. Recommended by Past Associate Editor, A. M. Annaswamy. This work was supported in part by the National Science Foundation under Grant ECS-9203491 and in part by the Air Force Office of Scientific Research under Grant F-49620-92-J-0495.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA.

IEEE Log Number 9408266.

The paper is organized as follows. After the problem statement in Section II, in Section III we design the ISS controllers and prove that the input-to-state stability is achieved. Section IV presents the nonlinear swapping lemma. Parameter identifiers with gradient and least-squares update laws are developed in Section V, and the stability proofs for the resulting adaptive systems are given in Section VI. In Section VII we analyze performance of the new adaptive systems. To reveal the connection with linear estimation-based designs we present in Section VIII the design of a weaker SG-controller. The new controller designs and performance are illustrated by examples in Section IX.

II. PROBLEM STATEMENT

The problem is to adaptively control nonlinear systems transformable into the parametric strict-feedback form

$$\begin{aligned} \dot{x}_i &= x_{i+1} + \theta^T \varphi_i(x_1, \dots, x_i), \quad 1 \leq i \leq n-1 \\ \dot{x}_n &= f_0(x)u + \theta^T \varphi_n(x) \\ y &= x_1 \end{aligned} \quad (2.1)$$

where $\theta \in \mathbb{R}^l$ is the vector of unknown constant parameters, f_0 and the components of $\phi = [\varphi_1 \dots \varphi_n]$ are smooth nonlinear functions in \mathbb{R}^n and $f_0(x) \neq 0 \forall x \in \mathbb{R}^n$. Necessary and sufficient conditions for a nonlinear system to be transformable into the form (2.1) are given in [12]. It should be noted that (2.1) is feedback linearizable for any bounded $\theta \in \mathbb{R}^p$.

The control objective is to force the output y of the system (2.1) to asymptotically track the output y_r of a known linear reference model while keeping all the closed-loop signals bounded. The reference model has the form

$$\begin{aligned} \dot{x}_m &= \begin{bmatrix} 0 & & \\ & I_{n-1} & \\ 0 & & \end{bmatrix} x_m + \begin{bmatrix} 0 \\ 0 \\ k_m \end{bmatrix} u_r \\ y_r &= x_{m,1} \end{aligned} \quad (2.2)$$

where $M(s) = s^n + m_{n-1}s^{n-1} + \dots + m_1s + m_0$ is Hurwitz, $k_m > 0$ and $u_r(t)$ is bounded and piecewise continuous. Another way of stating the same objective is to asymptotically track a given reference signal $y_r(t)$ with its first n derivatives known, bounded and piecewise continuous.

The above problem was first posed and solved in [12] using np estimates for p unknown parameters. This number of estimates was subsequently reduced in half in [8]. The over-parametrization was completely removed in [19] by the use of "tuning functions." In [40] the adaptive scheme of [12] was extended and recast in the observer based setting for the case when the nonlinearities in (2.1) are polynomial, a solution employing growth conditions was given in [33]. Possibilities to enlarge the class of systems that can be adaptively stabilized using the approach of [12] were explored in [1] and [36].

The Lyapunov-based results [12], [8], [19], and [36] employ only one type of parameter update laws. To increase the

flexibility in the update law selection we now develop an estimation-based design which treats the controller and identifier as separate modules.

Notation. For vectors we use $\|x\|_P \triangleq (x^T P x)^{1/2}$ to denote the weighted Euclidean norm of x . For matrices $\|X\|_F \triangleq (\text{tr}\{X^T X\})^{1/2} = (\text{tr}\{X X^T\})^{1/2}$ denote the Frobenius and $\|X\|_2$ the induced 2-norm of X . The \mathcal{L}_∞ , \mathcal{L}_2 and \mathcal{L} norms for signals are denoted by $\|\cdot\|_\infty$, $\|\cdot\|_2$ and $\|\cdot\|$ respectively. By referring to a matrix $A(t)$ as exponentially stable we mean that the corresponding LTV system $\dot{x} = A(t)x$ is exponentially stable. The spaces of all signals which are globally bounded, locally bounded and square integrable on $[0, t_f)$, $t_f > 0$ are denoted by $\mathcal{L}_\infty[0, t_f)$, $\mathcal{L}_\infty^l[0, t_f)$ and $\mathcal{L}_2[0, t_f)$ respectively. By saying that a signal belongs to $\mathcal{L}_\infty[0, t_f)$ or to $\mathcal{L}_2[0, t_f)$ we mean that the corresponding bound is independent of t_f .

III. ISS CONTROLLER DESIGN

Our modular estimation-based adaptive design for (2.1) places the burden of achieving boundedness on the controller module. We require that the controller guarantee input to state stability (ISS) with respect to the parameter error $\theta - \hat{\theta} = \tilde{\theta}$ and its derivative $\dot{\theta} = \dot{\tilde{\theta}}$ as disturbance inputs.

Using the backstepping procedure, which is well known from [12], [19] and [18] the adaptive nonlinear controller is recursively designed as follows:

$$z_i = x_i - x_{m,i} - \alpha_{i-1}$$

$$\begin{aligned} \alpha_i(\bar{x}_i, \tilde{\theta}, \bar{x}_{m,i-1}) &= -x_{i-1} - c_i z_i - \theta^T w_i \\ &+ \sum_{k=1}^{i-1} \left(\frac{\partial \alpha_{i-1}}{\partial x_k} x_{k+1} + \frac{\partial \alpha_{i-1}}{\partial x_{m,k}} x_{m,k+1} \right) \\ &- s_i(\bar{x}_i, \tilde{\theta}, \bar{x}_{m,i-1}) \end{aligned}$$

$$u_r(\bar{x}_i, \tilde{\theta}, \bar{x}_{m,i-1}) = \varphi_i - \sum_{k=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial x_k} \varphi_k, \quad i = 1, \dots, n$$

$$\begin{aligned} u &= \frac{1}{f_0(x)} [\alpha_n(\bar{x}_n, \tilde{\theta}, \bar{x}_{m,n-1}) - m_0 x_{m,1} - \\ &- m_{n-1} x_{m,n} + k_m u_r] \end{aligned} \quad (3.1)$$

where $\bar{x}_i = [x_1 \dots x_i]^T$, $\bar{x}_{m,i-1} = [x_{m,1} \dots x_{m,i-1}]^T$, $c_i > 0$, $i = 1, \dots, n$ and, for notational convenience $\alpha_0 \triangleq 0$, $\alpha_n \triangleq 0$. In these expressions the nonlinear damping functions $s_i(\bar{x}_i, \tilde{\theta}, \bar{x}_{m,i-1})$ are yet to be designed. We will employ these functions to achieve the desired ISS property of the system obtained by the recursive design procedure (3.1). This nonlinear system, called the error system, is readily shown to be

$$\dot{z} = A_-(z, \tilde{\theta}, t)z + W(z, \tilde{\theta}, t)^T \tilde{\theta} + D(z, \tilde{\theta}, t)^T \dot{\tilde{\theta}}, \quad z \in \mathbb{R}^n \quad (3.2)$$

where $z_1 = x_1 - x_{m1} = y - y_r$ represents the tracking error, and A_z, W, D are matrix-valued functions of z, θ and t

$$A_z(z, \theta, t) = \begin{bmatrix} -c_1 - \gamma_1 & 1 & 0 & 0 \\ -1 & -c_2 - \gamma_2 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 - c_n - \gamma_n \end{bmatrix}$$

$$W(z, \theta, t)^T = \begin{bmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_n^T \end{bmatrix} \in \mathbb{R}^{n \times p},$$

$$D(z, \theta, t)^T = \begin{bmatrix} 0 \\ -\frac{\partial \alpha_{i-1}}{\partial \theta} \\ \vdots \\ -\frac{\partial \alpha_{n-1}}{\partial \theta} \end{bmatrix} \in \mathbb{R}^{n \times p} \quad (3.3)$$

The explicit dependence of w_i and $\partial \alpha_{i-1} / \partial \theta$ (and hence γ_i) on t is due to the reference model, for example, $\varphi_1(t_1) = \varphi_1(z_1 + x_{m1}(t))$.

Except for the term $D(z, \theta, t)^T \theta$ the error system (3.2)–(3.3) is similar to the error system in [19] where the term $D(z, \theta, t)^T \theta$ was accounted for by using tuning functions. Here we let both θ and $\dot{\theta}$ appear as disturbance inputs. Their boundedness will later be guaranteed by parameter identifiers.

To design the nonlinear damping functions γ_i , we will employ the following lemma which evolved from [15] and [37].

Lemma 3.1 (Nonlinear Damping) Assume that for the system

$$\dot{x} = f(x, t) + g(x, t)[u + p(x, t)^T d(t)] \quad x \in \mathbb{R}^n, u \in \mathbb{R} \quad (3.4)$$

a feedback control $u = \mu(x, t)$ guarantees

$$\frac{\partial V}{\partial x} [f(x, t) + g(x, t)\mu(x, t)] + \frac{\partial V}{\partial t} \leq -U(x, t) \quad \forall x \in \mathbb{R}^n, \forall t \geq 0 \quad (3.5)$$

where $V, U: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ are positive definite and radially unbounded and V is decrescent and continuously differentiable in x uniformly in t ; $f: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}^n$, $g: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}^n$, $p: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}^q$, $\mu: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}$ are continuously differentiable in x and piecewise continuous and bounded in t , and $d: \mathbb{R}_+ \rightarrow \mathbb{R}^q$ is piecewise continuous. Then the feedback control

$$u = \mu(x, t) - \lambda |p(x, t)|^2 \frac{\partial V}{\partial x} (x, t) q(x, t) \quad (3.6)$$

where $\lambda > 0$, guarantees that

- 1) If $d \in \mathcal{L}_\infty$ then $x \in \mathcal{L}_\infty$.
- 2) If $d \in \mathcal{L}_2$ and $U(x, t) \geq c|x|^2 \quad \forall x \in \mathbb{R}^n, \forall t \geq 0$, $c > 0$ then $x \in \mathcal{L}_2$. If, in addition, $d \in \mathcal{L}_\infty$ then $\tau \in \mathcal{L}_\infty$ and $x(t) \rightarrow 0$ as $t \rightarrow \infty$.

Proof 1) Due to (3.5), the derivative of V along (3.4)–(3.6) is

$$\begin{aligned} \dot{V} &= \frac{\partial V}{\partial x} \left[f + g\mu + g \left(-\lambda p^T p \frac{\partial V}{\partial x} q + p^T d \right) \right] + \frac{\partial V}{\partial t} \\ &\leq -U - \lambda \left| p \frac{\partial V}{\partial x} q - \frac{1}{2\lambda} d \right|^2 + \frac{1}{4\lambda} |d|^2 \\ &\leq -U + \frac{1}{4\lambda} |d|^2 \end{aligned} \quad (3.7)$$

and, hence $x \in \mathcal{L}_\infty$.

2) Integrating (3.7) over $[0, \infty)$, we obtain

$$c \|x\|_2^2 \leq \|U\|_1 \leq \frac{1}{4\lambda} \|d\|_2^2 + V(0) \quad (3.8)$$

which implies that $x \in \mathcal{L}_2$. If, in addition, $d \in \mathcal{L}_\infty$ then by part 1) of this lemma, $x \in \mathcal{L}_\infty$ and therefore, $u \in \mathcal{L}_\infty$. Hence $x \in \mathcal{L}_\infty$. By Barbalat's lemma, $x(t) \rightarrow 0$ as $t \rightarrow \infty$. \square

To apply this lemma to the error system (3.2)–(3.3) we first note that the coefficients multiplying θ and $\dot{\theta}$ are w_i and $-\partial \alpha_{i-1} / \partial \theta$ respectively. They play the role of the function p in the lemma, while the part of $(\partial V / \partial x) q$ in (3.6) is played by z_i . Therefore, our choice of nonlinear damping functions is

$$\gamma_i = \kappa_i |w_i|^2 + q_i \left| \frac{\partial \alpha_{i-1}}{\partial \theta} \right|^2 \quad (3.9)$$

where $\kappa_i, q_i, i = 1, \dots, n$ are positive scalar constants.¹ The usefulness of the first term for achieving boundedness was stressed by Kanellakopoulos [16].

With this choice of γ_i , we now prove input-to-state stability of the error system (3.2), (3.3), (3.9) making use of the following constants $c_0 = \min_{1 \leq i \leq n} c_i$, $1/\kappa_0 = \sum_{i=1}^n (1/\kappa_i)$ and $1/q_0 = \sum_{i=1}^n (1/q_i)$.

Lemma 3.2 (ISS) In the error system (3.2), (3.3), (3.9) if $\theta, \dot{\theta} \in \mathcal{L}_\infty[0, t_f)$ then $z, x \in \mathcal{L}_\infty[0, t_f)$ and

$$|z(t)| \leq \frac{1}{2\sqrt{c_0}} \left(\frac{1}{\kappa_0} \|\theta\|_\infty^2 + \frac{1}{q_0} \|\dot{\theta}\|_\infty^2 \right)^{1/2} + |z(0)| e^{-c_0 t} \quad (3.10)$$

Proof Differentiating $\frac{1}{2}|z|^2$ along the solutions of (3.2) we compute

$$\begin{aligned} \frac{d}{dt} \left(\frac{1}{2} |z|^2 \right) &= - \sum_{i=1}^n c_i z_i^2 - \sum_{i=1}^n \left(\kappa_i |w_i|^2 + q_i \left| \frac{\partial \alpha_{i-1}}{\partial \theta} \right|^2 \right) z_i^2 \\ &\quad + \sum_{i=1}^n z_i \left(w_i^T \theta - \frac{\partial \alpha_{i-1}}{\partial \theta} \dot{\theta} \right) \\ &\leq -c_0 |z|^2 - \sum_{i=1}^n \kappa_i \left| w_i z_i - \frac{1}{2\kappa_i} \dot{\theta} \right|^2 \\ &\quad - \sum_{i=1}^n q_i \left| \frac{\partial \alpha_{i-1}}{\partial \theta} z_i + \frac{1}{2q_i} \dot{\theta} \right|^2 \\ &\quad + \left(\sum_{i=1}^n \frac{1}{4\kappa_i} \right) |\dot{\theta}|^2 + \left(\sum_{i=1}^n \frac{1}{4q_i} \right) |\dot{\theta}|^2 \end{aligned} \quad (3.11)$$

¹The constant coefficients q_i are not components of the vector field $g(x)$.

and arrive at

$$\frac{d}{dt} \left(\frac{1}{2} |z|^2 \right) \leq -c_0 |z|^2 + \frac{1}{4} \left(\frac{1}{\kappa_0} |\hat{\theta}|^2 + \frac{1}{g_0} |\dot{\hat{\theta}}|^2 \right). \quad (3.12)$$

From Lemma A.1(i), it follows that

$$\begin{aligned} |z(t)|^2 &\leq |z(0)|^2 e^{-2c_0 t} \\ &\quad + \frac{1}{2} \int_0^t e^{-2c_0(t-\tau)} \left(\frac{1}{\kappa_0} |\hat{\theta}(\tau)|^2 + \frac{1}{g_0} |\dot{\hat{\theta}}(\tau)|^2 \right) d\tau \\ &\leq |z(0)|^2 e^{-2c_0 t} + \frac{1}{4c_0} \left(\frac{1}{\kappa_0} \|\hat{\theta}\|_\infty^2 + \frac{1}{g_0} \|\dot{\hat{\theta}}\|_\infty^2 \right) \end{aligned} \quad (3.13)$$

which proves $z \in \mathcal{L}_\infty$ and (3.10), and by (3.1), $x \in \mathcal{L}_\infty$. \square

The quadratic form of the nonlinear damping functions is only one out of many possible forms. Any power greater than one would yield an ISS property, but the proof with quadratic nonlinear damping is by far the simplest.

A consequence of Lemma 3.2 is that, even when the adaptation is switched off, that is, when the parameter estimate $\hat{\theta}$ is constant ($\dot{\hat{\theta}} = 0$) and the only disturbance input is $\tilde{\theta}$, the state z of the error system (3.2), (3.3), (3.9) remain bounded and converges exponentially to a positively invariant compact set. (Note that since $\dot{\hat{\theta}} = 0$, the terms $-g_i |(\partial \alpha_{i-1} / \partial \hat{\theta})^T|^2 z_i$ are not needed.) Moreover, when the adaptation is switched off, this boundedness result holds even when the unknown parameter is time varying.

Corollary 3.1 (Boundedness Without Adaptation): If $\theta: \mathbb{R}_+ \rightarrow \mathbb{R}^p$ is piecewise continuous and bounded, and $\hat{\theta}$ is constant, then $z, x \in \mathcal{L}_\infty$, and

$$|z(t)| \leq \frac{1}{2\sqrt{c_0 \kappa_0}} \sup_{\tau \geq 0} |\theta(\tau) - \hat{\theta}| + |z(0)| e^{-c_0 t}. \quad (3.14)$$

Proof: Since $\dot{\hat{\theta}}(t) \equiv 0$, (3.12) holds with $\hat{\theta}(t) = \theta(t) - \hat{\theta}$. \square

Thus, the controller module alone guarantees boundedness, and the task of the adaptation is to achieve tracking.

IV. NONLINEAR SWAPPING

The desired boundedness property having been achieved by the controller module, we can now proceed to the identifier module design. To make this design as close to linear designs as possible, we derive a nonlinear counterpart of the ubiquitous Swapping Lemma [25]. This lemma is an analytical device which uses regressor filtering to account for the time-varying nature of the parameter estimates. It was used in the early nonlinear estimation-based results [27], [24], [29], [30], [35], [2], [9], [10], [39]. For a class of nonlinear systems, including our error system (3.2), we provide the following two nonlinear swapping lemmas.

Lemma 4.1 (Nonlinear Swapping): Consider the nonlinear time-varying system

$$\begin{aligned} \Sigma_1: \quad \dot{z} &= A(z, t)z + g(z, t)W(z, t)^T \hat{\theta} - D(z, t)^T \dot{\hat{\theta}} \\ y_1 &= h(z, t)z + l(z, t)W(z, t)^T \hat{\theta} \end{aligned} \quad (4.1)$$

where $\hat{\theta}: \mathbb{R}_+ \rightarrow \mathbb{R}^p$ is differentiable, $A: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}^{n \times n}$, $g: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}^{n \times m}$, $W: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}^{p \times n}$, $D: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}^{p \times n}$, $l: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}^{r \times m}$ are locally Lipschitz in z and continuous and bounded in t , and $h: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}^{r \times n}$ is bounded in z and t . Along with (4.1) consider the linear time-varying systems

$$\Sigma_2: \quad \begin{aligned} \dot{\chi}^T &= A(z, t)\chi^T + g(z, t)W(z, t)^T \\ y_2 &= h(z, t)\chi^T + l(z, t)W(z, t)^T \end{aligned} \quad (4.2)$$

$$\Sigma_3: \quad \begin{aligned} \dot{\psi} &= A(z, t)\psi + \chi^T \dot{\hat{\theta}} + D(z, t)^T \dot{\hat{\theta}} \\ y_3 &= -h(z, t)\psi. \end{aligned} \quad (4.3)$$

Assume that $z(t)$ is continuous on $[0, \infty)$ and there exists a continuously differentiable function $V: \mathbb{R}^n \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that

$$\alpha_1 |\zeta|^2 \leq V(\zeta, t) \leq \alpha_2 |\zeta|^2 \quad (4.4)$$

and for each $z \in C^0$

$$\frac{\partial V}{\partial \zeta} A(z, t)\zeta + \frac{\partial V}{\partial t} \leq -\alpha_3 |\zeta|^2 \quad (4.5)$$

$\forall t \geq 0, \forall \zeta \in \mathbb{R}^n, \alpha_1, \alpha_2, \alpha_3 > 0$. Then for $\forall z(0), \psi(0) \in \mathbb{R}^n, \forall \chi(0) \in \mathbb{R}^{p \times n}, \forall t \geq 0$ the outputs of systems (4.1)–(4.3) are related by

$$y_1 = y_2 \hat{\theta} + y_3 + y_e \quad (4.6)$$

where y_e is bounded and exponentially decaying.

Proof: Due to the continuity of $z(t)$, we see that $g(z(t), t)$, $W(z(t), t)$ and $D(z(t), t)$ are continuous in t . Since $gW \in \mathcal{L}_{\infty}$ and Σ_2 is a linear time-varying system, then $\chi \in \mathcal{L}_{\infty}$. Therefore $(\chi + D)^T \hat{\theta} \in \mathcal{L}_{\infty}$, which implies $\psi \in \mathcal{L}_{\infty}$ because Σ_3 is a linear time-varying system. Differentiating $\tilde{z} = z + \psi - \chi^T \hat{\theta}$, we obtain

$$\dot{\tilde{z}} = \dot{z} + \dot{\psi} - \dot{\chi}^T \hat{\theta} - \chi^T \dot{\hat{\theta}} = A(z, t)\tilde{z} \quad (4.7)$$

which together with (4.4)–(4.5) yields

$$\dot{V}(\tilde{z}, t) = \frac{\partial V}{\partial \tilde{z}} A(z, t)\tilde{z} + \frac{\partial V}{\partial t} \leq -\alpha_3 |\tilde{z}|^2 \leq -\frac{\alpha_3}{\alpha_2} V. \quad (4.8)$$

Therefore $V(t) \leq V(0)e^{-(\alpha_3/\alpha_2)t}$, and, hence

$$|\tilde{z}(t)| \leq \sqrt{\frac{\alpha_2}{\alpha_1}} |\tilde{z}(0)| e^{-(\alpha_3/2\alpha_2)t}. \quad (4.9)$$

Now, (4.1)–(4.3) imply that $y_e = y_1 - y_2 \hat{\theta} - y_3 = h(z, t)\tilde{z}$. Since $h(z, t)$ is bounded then y_e is bounded and decays to zero exponentially.

Remark 4.1: When $D(z, t) \equiv 0$, the result of Lemma 4.1 is reminiscent of Morse's linear Swapping Lemma [25]. To see this we rewrite (4.6) as

$$T_z[W^T \hat{\theta}] = T[W^T] \hat{\theta} + T_h[T_g[W^T] \hat{\theta}] + y_e. \quad (4.10)$$

In this notation $T_z W^T \tilde{\theta} \mapsto y_1$ is the nonlinear operator defined by (4.1) with $D(z, t) \equiv 0$, while the system

$$\begin{aligned}\xi &= A(z(t), t)\xi + g(z(t), t)u \\ y &= h(z(t), t)\xi + l(z(t), t)u\end{aligned}\quad (4.11)$$

is used to define the linear time-varying operators $T u \mapsto y$, $T_g u \mapsto y$ for $h = I$ and $l = 0$, $T_h u \mapsto y$ for $g = I$ and $l = 0$. When A , g , h and l are constant, then the operator $T_z(y) = T(s) = h(sI - A)^{-1}g + l$ is a proper stable rational transfer function, $T_g(s) = (sI - A)^{-1}g$, $T_h(s) = -h(sI - A)^{-1}$, and Lemma 4.1 reduces to Lemma 3.6.5 from [34]. \square

In some texts on adaptive linear control, an extended result which guarantees that $\tilde{\theta} \in \mathcal{L}_2 \Rightarrow T_z[W^T \tilde{\theta}] - T[W^T] \tilde{\theta} \in \mathcal{L}_2$ is also referred to as Swapping Lemma. Our next lemma is a nonlinear time-varying generalization of this result.

Lemma 4.2 Consider systems (4.1)–(4.3) with the same set of assumptions as in Lemma 4.1. Further, assume that $z \in \mathcal{L}_\infty$. If $\tilde{\theta} \in \mathcal{L}_2$, then

$$y_1 - y_2 \tilde{\theta} \in \mathcal{L}_2 \quad (4.12)$$

If $\theta \in \mathcal{L}_2 \cap \mathcal{L}_\infty$ then

$$\lim_{t \rightarrow \infty} [y_1(t) - y_2(t)\theta(t)] = 0 \quad (4.13)$$

Proof Since $z \in \mathcal{L}_\infty$ then $gW^T D \in \mathcal{L}_\infty$. Due to the exponential stability of $A(z, t)$, it follows that $\chi \in \mathcal{L}_\infty$. By Lemma 4.1, $y_e \in \mathcal{L}_2$. We need to prove that $y_3 \in \mathcal{L}_2$. The solution of (4.3) is

$$\begin{aligned}\psi(t) &= \Phi_z(t, 0)\psi(0) \\ &+ \int_0^t \Phi_z(t, \tau)[\chi(\tau) + D(z(\tau), \tau)]^T \tilde{\theta}(\tau) d\tau\end{aligned}\quad (4.14)$$

where (4.4)–(4.5) guarantee that the state transition matrix $\Phi_z: \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}^{n \times n}$ is such that $|\Phi_z(t, \tau)|_2 \leq ke^{-\alpha(t-\tau)}$, $k, \alpha > 0$. Since χ and D are bounded then

$$\begin{aligned}|\psi(t)| &\leq ke^{-\alpha t}|\psi(0)| + k\|\chi + D\|_\infty \int_0^t e^{-\alpha(t-\tau)}|\tilde{\theta}(\tau)| d\tau \\ &\leq ke^{-\alpha t}|\psi(0)| + k\|\chi + D\|_\infty \left(\int_0^t e^{-\alpha(t-\tau)} d\tau \right)^{1/2} \\ &\quad \left(\int_0^t e^{-\alpha(t-\tau)} |\tilde{\theta}(\tau)|^2 d\tau \right)^{1/2} \\ &\leq ke^{-\alpha t}|\psi(0)| + k\|\chi + D\|_\infty \frac{1}{\sqrt{\alpha}} \\ &\quad \left(\int_0^t e^{-\alpha(t-\tau)} |\tilde{\theta}(\tau)|^2 d\tau \right)^{1/2}\end{aligned}\quad (4.15)$$

where the second inequality is obtained using the Schwartz inequality. By squaring (4.15) and integrating over $[0, t]$ we obtain

$$\begin{aligned}\int_0^t |\psi(\tau)|^2 d\tau &\leq \frac{k^2}{2\alpha} |\psi(0)|^2 + \frac{k^2}{\alpha} \|\chi + D\|_\infty^2 \\ &\quad \int_0^t \left[\int_0^\tau e^{-\alpha(\tau-s)} |\tilde{\theta}(s)|^2 ds \right] d\tau\end{aligned}\quad (4.16)$$

Changing the sequence of integration, (4.16) becomes

$$\begin{aligned}\int_0^t |\psi(\tau)|^2 d\tau &\leq \frac{k^2}{2\alpha} |\psi(0)|^2 + \frac{k^2}{\alpha} \|\chi + D\|_\infty^2 \\ &\quad \int_0^t e^{-\alpha s} |\tilde{\theta}(s)|^2 \left(\int_s^t e^{-\alpha \tau} d\tau \right) ds \\ &\leq \frac{k^2}{2\alpha} |\psi(0)|^2 + \frac{k^2}{\alpha} \|\chi + D\|_\infty^2 \\ &\quad \int_0^t e^{-\alpha s} |\tilde{\theta}(s)|^2 \frac{1}{\alpha} e^{-\alpha s} ds\end{aligned}\quad (4.17)$$

because $\int_s^t e^{-\alpha \tau} d\tau = 1/\alpha (e^{-\alpha s} - e^{-\alpha t}) \leq (1/\alpha) e^{-\alpha s}$. Now the cancellation $e^{\alpha s} e^{-\alpha s} = 1$ in (4.17) yields

$$\|\psi\|_2 \leq \frac{k}{\sqrt{2\alpha}} |\psi(0)| + \frac{k}{\alpha} \|\chi + D\|_\infty \|\tilde{\theta}\|_2 < \infty \quad (4.18)$$

which proves $\psi \in \mathcal{L}_2$. Due to the uniform boundedness of h it follows that $y_3 \in \mathcal{L}_2$. This proves (4.12). When $\theta \in \mathcal{L}_2 \cap \mathcal{L}_\infty$ then $\psi \in \mathcal{L}_2 \cap \mathcal{L}_\infty$ and $\psi \in \mathcal{L}_\infty$. Thus, by Barbalat's lemma $\psi(t) \rightarrow 0$ and hence $y_3(t) \rightarrow 0$ as $t \rightarrow \infty$. This proves (4.13) because $y_e(t) \rightarrow 0$ as $t \rightarrow \infty$.

Remark 4.2 When $D(\cdot, t) \equiv 0$ we rewrite (4.12) as

$$T[W^T \tilde{\theta}] - I[W^T] \tilde{\theta} \in \mathcal{L}_2 \quad (4.19)$$

and (4.13) as

$$\lim_{t \rightarrow \infty} \{T_z[W^T \tilde{\theta}](t) - (I[W^T] \tilde{\theta})(t)\} = 0 \quad (4.20)$$

with T_z and T as in Remark 4.1. For constant A , g , h and l the operator $T_z = T$ is a proper stable rational transfer function, and Lemma 4.2 reduces to Lemma 2.11 from [28]. \square

V. PARAMETER IDENTIFIERS

We are now in the position to design the identifier module by applying the Nonlinear Swapping Lemma 4.1 to either z - or x -system. Each of the two types of identifiers, with z -swapping and with x -swapping, can be implemented with either gradient or least-squares update laws. These parameter identifiers are variants of the regressor filtering identifiers in [32].

A. z -Swapping

For the error system (3.2) we introduce the filters

$$\begin{aligned}\chi_0 &= A_z(z, \theta, t)\chi_0 + W(z, \theta, t)^T \theta - D(z, \theta, t)^T \theta, \\ \chi_0 &\in \mathbb{R}^n\end{aligned}\quad (5.1)$$

$$\chi^T = A_z(z, \theta, t)\chi^T + W(z, \theta, t)^T, \quad \chi \in \mathbb{R}^{p \times n} \quad (5.2)$$

and define the estimation error as

$$\epsilon = z + \chi_0 - \chi^T \theta, \quad \epsilon \in \mathbb{R}^n \quad (5.3)$$

Along with ϵ we define

$$\epsilon = z + \chi_0 - \chi^T \theta \quad \epsilon \in \mathbb{R}^n \quad (5.4)$$

Then we obtain

$$\dot{\epsilon} = \chi^T \dot{\theta} + \dot{\epsilon} \quad (5.5)$$

and, by differentiating (5.4) and substituting (3.2), (5.1) and (5.2), recognize that \tilde{e} is governed by

$$\dot{\tilde{e}} = A_z(z, \hat{\theta}, t)\tilde{e}. \quad (5.6)$$

The update laws for $\hat{\theta}$ employ the estimation error ϵ and the filtered regressor χ . The gradient update law is

$$\dot{\hat{\theta}} = \Gamma \frac{\chi \epsilon}{1 + \nu |\chi|_{\mathcal{F}}^2}, \quad \Gamma = \Gamma^T > 0, \nu \geq 0 \quad (5.7)$$

and the least-squares law is

$$\dot{\hat{\theta}} = \Gamma \frac{\chi \epsilon}{1 + \nu |\chi|_{\mathcal{F}}^2}$$

$$\dot{\Gamma} = -\Gamma \frac{\chi \chi^T}{1 + \nu |\chi|_{\mathcal{F}}^2} \Gamma, \quad \Gamma(0) = \Gamma^T(0) > 0, \nu \geq 0. \quad (5.8)$$

By allowing $\nu = 0$ we encompass unnormalized update laws.

Since the regressor χ is a matrix, we use the Frobenius norm $|\chi|_{\mathcal{F}}$ to avoid the need for on-line matrix inversion, as well as unnecessary algebraic complications in the stability arguments that would arise from applying update laws $\dot{\hat{\theta}} = \Gamma \chi (I_p + \nu \chi^T \Gamma \chi)^{-1} \epsilon$ with Γ fixed or updated with $\dot{\Gamma} = -\Gamma \chi (I_p + \nu \chi^T \Gamma \chi)^{-1} \chi^T \Gamma$.

The boundedness properties of the z -swapping identifiers are as follows.

Lemma 5.1: Suppose the solution $x(t)$ is defined on $[0, t_f)$. The update laws (5.7) and (5.8) guarantee that

- 1) if $\nu = 0$ then $\hat{\theta} \in \mathcal{L}_{\infty}[0, t_f)$ and $\epsilon \in \mathcal{L}_2[0, t_f)$,
- 2) if $\nu > 0$ then $\hat{\theta} \in \mathcal{L}_{\infty}[0, t_f)$ and

$$\hat{\theta}, \frac{\epsilon}{\sqrt{1 + \nu |\chi|_{\mathcal{F}}^2}} \in \mathcal{L}_2[0, t_f) \cap \mathcal{L}_{\infty}[0, t_f).$$

Proof: (Sketch) Noting from (5.6) and (3.3) that $d/dt (\frac{1}{2} |\tilde{e}|^2) = -\sum_{i=1}^n c_i \tilde{e}_i^2 \leq -c_0 |\tilde{e}|^2$ it is clear that the positive definite function $V = \frac{1}{2} |\hat{\theta}|_{\Gamma^{-1}}^2 + 1/2 c_0 |\tilde{e}|^2$ can be used as in [6], [34], [7] to prove the lemma. \square

As explained in [6], various modifications of the least-squares algorithm (covariance resetting, exponential data weighting, etc.) do not affect the properties established by Lemma 5.1. *A priori* knowledge of parameter bounds can also be incorporated via projection.

B. x -Swapping

A different identifier results if instead of the error system (3.2) we consider the plant (2.1) rewritten in the form

$$\dot{x} = Ex + e_n \beta_0(x)u + \phi(x)^T \theta \quad (5.9)$$

where

$$E = \begin{bmatrix} 0 & & & \\ \vdots & I_{n-1} & & \\ 0 & \cdots & 0 & \end{bmatrix}.$$

We employ the following filters

$$\dot{\Omega}_0 = \bar{A}(t)(\Omega_0 - x) + Ex + e_n \beta_0(x)u, \quad \Omega_0 \in \mathbb{R}^n \quad (5.10)$$

$$\dot{\Omega}^T = \bar{A}(t)\Omega^T + \phi(x)^T, \quad \Omega \in \mathbb{R}^{p \times n} \quad (5.11)$$

where $\bar{A}(t)$ is an exponentially stable matrix. We define the estimation error vector

$$\epsilon = x - \Omega_0 - \Omega^T \hat{\theta}, \quad \epsilon \in \mathbb{R}^n. \quad (5.12)$$

and along with it

$$\tilde{\epsilon} = x - \Omega_0 - \Omega^T \theta, \quad \tilde{\epsilon} \in \mathbb{R}^n. \quad (5.13)$$

Then we obtain

$$\epsilon = \Omega^T \tilde{\theta} + \tilde{\epsilon} \quad (5.14)$$

and, by differentiating (5.13) and substituting (5.9), (5.10) and (5.11), recognize that $\tilde{\epsilon}$ is governed by

$$\dot{\tilde{\epsilon}} = \bar{A}(t)\tilde{\epsilon}. \quad (5.15)$$

The update laws for $\hat{\theta}$ employ the estimation error ϵ and the filtered regressor Ω . The gradient update law is

$$\dot{\hat{\theta}} = \Gamma \frac{\Omega \epsilon}{1 + \nu |\Omega|_{\mathcal{F}}^2}, \quad \Gamma = \Gamma^T > 0, \nu \geq 0 \quad (5.16)$$

and the least-squares law is

$$\dot{\hat{\theta}} = \Gamma \frac{\Omega \epsilon}{1 + \nu |\Omega|_{\mathcal{F}}^2}$$

$$\dot{\Gamma} = -\Gamma \frac{\Omega \Omega^T}{1 + \nu |\Omega|_{\mathcal{F}}^2} \Gamma, \quad \Gamma(0) = \Gamma^T(0) > 0, \nu \geq 0. \quad (5.17)$$

Again, by allowing $\nu = 0$ we encompass unnormalized gradient and least-squares. Concerning the update law modifications, the same comments from the preceding subsection are also in order here.

Lemma 5.2: Suppose $x(t)$ is defined on $[0, t_f)$, and $\bar{A}(t)$ is continuous and bounded on $[0, t_f)$ and exponentially stable. The update laws (5.16) and (5.17) guarantee that

- 1) if $\nu = 0$ then $\hat{\theta} \in \mathcal{L}_{\infty}[0, t_f)$ and $\epsilon \in \mathcal{L}_2[0, t_f)$,
- 2) if $\nu > 0$ then $\hat{\theta} \in \mathcal{L}_{\infty}[0, t_f)$ and

$$\hat{\theta}, \frac{\epsilon}{\sqrt{1 + \nu |\Omega|_{\mathcal{F}}^2}} \in \mathcal{L}_2[0, t_f) \cap \mathcal{L}_{\infty}[0, t_f).$$

Proof: (Sketch) There exists a continuously differentiable, bounded, positive definite, symmetric $P: \mathbb{R}_+ \rightarrow \mathbb{R}^{n \times n}$ such that $\dot{P} + P\bar{A} + \bar{A}^T P = -I, \forall t \in [0, t_f)$, and the positive definite function $V = \frac{1}{2} \|\hat{\theta}\|_{P^{-1}}^2 + |\tilde{\epsilon}|_P^2$ can be used as in [6], [34], [7] to prove the lemma. \square

VI. STABILITY AND TRACKING

Either of the identifiers from the preceding sections can now be connected with the ISS-controller (3.1), (3.9). We give stability proofs for the resulting adaptive systems. These proofs encompass both normalized and unnormalized update laws.

Theorem 6.1 (z -Swapping Scheme): All the signals in the adaptive system consisting of the plant (2.1), controller (3.1), (3.9), filters (5.1), (5.2), and either the gradient (5.7) or the least-squares (5.8) update law, are globally uniformly bounded for all $t \geq 0$, and $\lim_{t \rightarrow \infty} z(t) = 0$. This means, in particular, that global asymptotic tracking is achieved: $\lim_{t \rightarrow \infty} [y(t) - y_r(t)] = 0$. Furthermore, if $\lim_{t \rightarrow \infty} r(t) = 0$ and $\phi(0) = 0$ then $\lim_{t \rightarrow \infty} x(t) = 0$.

Proof: Due to the continuity of x_m and the smoothness of the nonlinear terms appearing in (2.1), (3.1), (3.9), (5.1), (5.2), (5.7), (5.8), the solution of the closed-loop adaptive system exists and is unique. Let its maximum interval of existence be $[0, t_f)$.

For the normalized update laws, from Lemma 5.1 we obtain

$$\bar{\theta}, \dot{\bar{\theta}}, \frac{\epsilon}{\sqrt{1+\nu|\chi|^2_{\mathcal{F}}}} \in \mathcal{L}_{\infty}[0, t_f).$$

When the update laws are unnormalized, Lemma 5.1 gives only $\bar{\theta} \in \mathcal{L}_{\infty}[0, t_f)$ and we have to establish boundedness of $\dot{\bar{\theta}}$. To this end, we treat (5.2) in a fashion similar to (3.11)

$$\begin{aligned} & \frac{d}{dt} \left(\frac{1}{2} |\chi|_{\mathcal{F}}^2 \right) \\ &= \text{tr} \frac{d}{dt} \left(\frac{1}{2} \chi \chi^T \right) \\ &\leq -c_0 \text{tr} \{ \chi \chi^T \} - \text{tr} \{ \chi \text{diag}(\kappa_1 |w_1|^2, \dots, \kappa_n |w_n|^2) \chi^T \} \\ &\quad + \text{tr} \{ W^T \chi \} \\ &= -c_0 |\chi|_{\mathcal{F}}^2 + \sum_{i=1}^n (-\kappa_i |w_i|^2 |\chi_i|^2 + w_i^T \chi_i) \\ &\leq -c_0 |\chi|_{\mathcal{F}}^2 + \frac{1}{4\kappa_0}. \end{aligned} \quad (6.1)$$

This proves that $\chi \in \mathcal{L}_{\infty}[0, t_f)$. Therefore, by (5.5) and because of the boundedness of $\bar{\epsilon}$ we conclude that $\epsilon \in \mathcal{L}_{\infty}[0, t_f)$. Now by (5.7) or (5.8), $\dot{\bar{\theta}} \in \mathcal{L}_{\infty}[0, t_f)$. Therefore, by Lemma 3.2, $z, r \in \mathcal{L}_{\infty}[0, t_f)$. Finally, by (5.3), $\chi_0 \in \mathcal{L}_{\infty}[0, t_f)$.

We have thus shown that all of the signals of the closed-loop adaptive system are bounded on $[0, t_f)$ by constants depending only on the initial conditions, design gains, the external signals x_m and r , and not depending on t_f . The independence of the bounds of t_f proves that $t_f = \infty$. Hence, all signals are globally uniformly bounded on $[0, \infty)$.

Now we set out to prove that $z \in \mathcal{L}_2$, and eventually that $z(t) \rightarrow 0$ as $t \rightarrow \infty$. For the normalized update laws, from Lemma 5.1 we obtain $\dot{\bar{\theta}} \in \mathcal{L}_2$. Since $\chi \in \mathcal{L}_{\infty}$ then $\epsilon \in \mathcal{L}_2$. When the update laws are unnormalized Lemma 5.1 gives $\epsilon \in \mathcal{L}_2$, and since $\chi \in \mathcal{L}_{\infty}$ then by (5.7) or (5.8), $\dot{\bar{\theta}} \in \mathcal{L}_2$. Consequently in both the normalized and the unnormalized cases $\chi^T \dot{\bar{\theta}} \in \mathcal{L}_2$ because $\bar{\epsilon} \in \mathcal{L}_2$. With $V = \frac{1}{2} |\zeta|^2$, all the conditions of Lemmas 4.1 and 4.2 are satisfied. Thus, by Lemma 4.2, $z - \chi^T \dot{\bar{\theta}} \in \mathcal{L}_2$. Hence $z \in \mathcal{L}_2$. To prove the convergence of z to zero, we note that (3.2), (3.3) implies that $\dot{z} \in \mathcal{L}_{\infty}$. Therefore, by Barbalat's lemma $z(t) \rightarrow 0$ as $t \rightarrow \infty$. When $r(t) \rightarrow 0$ then $x_m(t) \rightarrow 0$ as $t \rightarrow \infty$, and from the definitions in (3.1) we conclude that, if $\phi(0) = 0$, then $x(t) \rightarrow 0$ as $t \rightarrow \infty$. \square

Now we proceed to prove stability of the x -swapping scheme. With normalized update laws, the proof is similar to the proof of Theorem 6.1. With the unnormalized update laws, it is not clear how to prove boundedness of all signals for

an arbitrary exponentially stable $\bar{A}(t)$. We avoid this difficulty by designing

$$\bar{A}(t) = A_0 - \lambda \phi^T(x) \phi(x) P \quad (6.2)$$

where $\lambda > 0$ and A_0 is an arbitrary constant matrix that satisfies $PA_0 + A_0^T P = -I$, $P = P^T > 0$. With this design the matrix $\bar{A}(t)$ is exponentially stable because

$$P\bar{A}(t) + \bar{A}^T(t)P = -I - 2\lambda P\phi^T\phi P \leq -I \quad (6.3)$$

Theorem 6.2 (x -Swapping Scheme) All the signals in the adaptive system consisting of the plant (2.1), controller (3.1), (3.9), filters (5.10), (5.11), and either the gradient (5.16) or the least-squares (5.17) update law are globally uniformly bounded for all $t \geq 0$, and $\lim_{t \rightarrow \infty} z(t) = 0$. This means, in particular, that global asymptotic tracking is achieved: $\lim_{t \rightarrow \infty} [y(t) - y_r(t)] = 0$. Furthermore, if $\lim_{t \rightarrow \infty} r(t) = 0$ and $\phi(0) = 0$ then $\lim_{t \rightarrow \infty} x(t) = 0$.

Proof We first consider the normalized update laws.

As in the proof of Theorem 6.1, we show that $\theta, \dot{\theta}, z, r \in \mathcal{L}_{\infty}[0, t_f)$ and hence $u \in \mathcal{L}_{\infty}[0, t_f)$. From (5.10) and (5.11) it follows that Ω_0, Ω , and therefore ϵ are in $\mathcal{L}_{\infty}[0, t_f)$. Now, by the same argument as in the proof of Theorem 6.1 we conclude that $t_f = \infty$.

Second, we consider the unnormalized update laws (5.16) and (5.17) with $\bar{A}(t)$ given by (6.2). Along the solutions of (5.11) we have

$$\begin{aligned} \frac{d}{dt} (\Omega P \Omega^T) &= -\Omega \Omega^T - 2\lambda \Omega P \phi^T \phi P \Omega^T + \Omega P \phi^T + \phi P \Omega^T \\ &= -\Omega \Omega^T - 2\lambda \left(\phi P \Omega^T - \frac{1}{2\lambda} I_p \right)^T \\ &\quad \cdot \left(\phi P \Omega^T - \frac{1}{2\lambda} I_p \right) + \frac{1}{2\lambda} I_p \end{aligned} \quad (6.4)$$

which implies

$$\frac{d}{dt} (\text{tr} \{ \Omega P \Omega^T \}) \leq -\text{tr} \{ \Omega \Omega^T \} + \frac{p}{2\lambda}. \quad (6.5)$$

Hence $\Omega \in \mathcal{L}_{\infty}[0, t_f)$. Lemma 5.2 gives² $\dot{\bar{\theta}} \in \mathcal{L}_{\infty}[0, t_f)$, and from (5.14) and (5.15) we conclude that $\epsilon \in \mathcal{L}_{\infty}[0, t_f)$. Now by (5.16) or (5.17), $\dot{\bar{\theta}} \in \mathcal{L}_{\infty}[0, t_f)$. Therefore, by Lemma 3.2, $z, r \in \mathcal{L}_{\infty}[0, t_f)$. Finally, by (5.12), $\Omega_0 \in \mathcal{L}_{\infty}[0, t_f)$. As before, $t_f = \infty$.

Now we set out to prove that $z \in \mathcal{L}_2$. For normalized update laws, from Lemma 5.2, we have that $\dot{\bar{\theta}}, \epsilon / \sqrt{1+\nu|\Omega|_{\mathcal{F}}^2} \in \mathcal{L}_2$. Since $\Omega \in \mathcal{L}_{\infty}$ then $\epsilon \in \mathcal{L}_2$. When the update laws are unnormalized, Lemma 5.2 gives $\epsilon \in \mathcal{L}_2$, and since $\Omega \in \mathcal{L}_{\infty}$ then by (5.16) or (5.17), $\dot{\bar{\theta}} \in \mathcal{L}_2$. Consequently for both the normalized and the unnormalized cases, $\Omega^T \dot{\bar{\theta}} \in \mathcal{L}_2$ because $\bar{\epsilon} \in \mathcal{L}_2$. Now, as in Theorem 6.1, we invoke Lemma 4.2 to deduce that $z - \chi^T \dot{\bar{\theta}} \in \mathcal{L}_2$. To show that $z \in \mathcal{L}_2$, we need to prove that $\Omega^T \dot{\bar{\theta}} \in \mathcal{L}_2$ implies $\chi^T \dot{\bar{\theta}} \in \mathcal{L}_2$, or, in the notation of Lemma A.2 from the Appendix, that $T_{\bar{A}}[\phi^T] \dot{\bar{\theta}} \in \mathcal{L}_2$ implies

²Since $\bar{A}(t)$ depends on $r(t)$ whose boundedness is yet to be proven, in invoking Lemma 5.2 we violate the boundedness condition for $\bar{A}(t)$. This, however, causes no difficulty because the boundedness condition is required only in order to establish the existence of P , and we know P in (6.2) *a priori*.

Let $[W^T]\tilde{\theta} \in \mathcal{L}_2$. To apply this lemma to our adaptive system note from (3.3) and (3.1) that

$$W^T(z, \hat{\theta}, t) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -\frac{\partial \alpha_1}{\partial x_1} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial \alpha_{n-1}}{\partial x_1} & -\frac{\partial \alpha_{n-1}}{\partial x_2} & \cdots & 1 \end{bmatrix} \phi^T(x) \triangleq M(z, \hat{\theta}, t) \phi^T(x). \quad (6.6)$$

Since $M(z(t), \hat{\theta}(t), t)$ satisfies the conditions of Lemma A.2 then $\chi^T \tilde{\theta} \in \mathcal{L}_2$ and hence $z \in \mathcal{L}_2$. The rest of the proof is the same as for Theorem 6.1. \square

Remark 6.1: All the above results are presented for the parametric-strict-feedback form (2.1) without zero dynamics. As in [12], they can be readily modified for the strict-feedback systems with zero-dynamics

$$\begin{aligned} \dot{x}_i &= x_{i+1} + \theta^T \varphi_i(x_1, \dots, x_i, x^r), & 1 \leq i \leq n-1 \\ \dot{x}_n &= \beta_0(x)u + \theta^T \varphi_n(x) \\ \dot{x}^r &= \Phi_0(y, x^r) + \Phi(y, x^r)\theta \\ y &= x_1 \end{aligned} \quad (6.7)$$

where the x^r -subsystem has a bounded-input bounded-state (BIBS) property with respect to y as its input. The procedure can also be modified, as in [12], to obtain a local result for the parametric-pure-feedback systems, i.e., the systems in which φ_i also depends on x_{i+1} . As in [40], the subset of pure-feedback systems that can be controlled globally can be enlarged using an appropriate filter and parameter estimate initialization. \square

VII. \mathcal{L}_∞ , MEAN-SQUARE AND \mathcal{L}_2 PERFORMANCE

For linear systems the issue of transient performance has recently received considerable attention (see [4], [20] and references therein). For the adaptive schemes presented in the preceding sections we now derive \mathcal{L}_∞ , mean-square, and \mathcal{L}_2 bounds for the error state z , which incorporate the bounds for the tracking error $y - y_r$.

First we give performance bounds for parameter identifiers and use them to establish \mathcal{L}_∞ and mean-square bounds for z that are valid for both the z -swapping and the x -swapping schemes. Then we derive an \mathcal{L}_2 norm bound on z for the z -swapping scheme. For the x -swapping scheme a similar \mathcal{L}_2 bound is not yet available.

We analyze in detail the scheme with the normalized gradient update laws and suggest in Remarks 7.1 and 7.4 how to modify the derivations for other update laws.

Without loss of generality we assume in our analysis, and recommend for implementation, that $\tilde{\epsilon}(0)$, $\chi(0)$ (in the z -swapping scheme), and $\Omega(0)$ (in the x -swapping scheme), be set to zero. This can be achieved by initializing $\chi_0(0) = -z(0)$, $\chi(0) = 0$, in the z -swapping scheme, and $\Omega_0(0) = x(0)$, $\Omega(0) = 0$, in the x -swapping scheme. For simplicity, we also let $\Gamma = \gamma I$. We explain in Remark 7.3 how the performance bounds differ in the absence of initialization.

Lemma 7.1: For both the z -swapping (5.7) and the x -swapping (5.16) normalized ($\nu > 0$) gradient update laws the following bounds hold

$$i) \quad \|\tilde{\theta}\|_\infty = |\tilde{\theta}(0)| \quad (7.1)$$

$$ii) \quad \|\dot{\tilde{\theta}}\|_\infty \leq \frac{\gamma}{\nu} |\tilde{\theta}(0)| \quad (7.2)$$

$$iii) \quad \|\dot{\tilde{\theta}}\|_2 \leq \sqrt{\frac{\gamma}{2\nu}} |\tilde{\theta}(0)|. \quad (7.3)$$

Proof: The proof is given for the z -swapping identifier (5.1), (5.2), (5.7). The proof for the x -swapping identifier (5.10), (5.11), (5.16) is identical.

Consider the positive definite function $V_{\tilde{\theta}} = 1/2\gamma|\tilde{\theta}|^2$. Its derivative along the solutions of (5.5), (5.7) is

$$\dot{V}_{\tilde{\theta}} = -\frac{|\epsilon|^2}{1 + \nu|\chi|_{\mathcal{F}}^2} \leq 0. \quad (7.4)$$

i) Due to the nonpositivity of $\dot{V}_{\tilde{\theta}}$ we have $V_{\tilde{\theta}}(t) \leq V_{\tilde{\theta}}(0)$ which implies (7.1).

ii) From (5.7) we can write

$$\begin{aligned} |\dot{\tilde{\theta}}|^2 &\leq \gamma^2 \frac{\epsilon^T \chi \chi^T \epsilon}{(1 + \nu|\chi|_{\mathcal{F}}^2)^2} \leq \gamma^2 \frac{|\epsilon|^2 |\chi|_{\mathcal{F}}^2}{(1 + \nu|\chi|_{\mathcal{F}}^2)^2} \\ &\leq \frac{\gamma^2}{\nu} \frac{|\epsilon|^2}{1 + \nu|\chi|_{\mathcal{F}}^2} \end{aligned} \quad (7.5)$$

By using (5.5) we get

$$|\dot{\tilde{\theta}}|^2 \leq \frac{\gamma^2}{\nu} \frac{\tilde{\theta}^T \chi \chi^T \tilde{\theta}}{1 + \nu|\chi|_{\mathcal{F}}^2} \leq \frac{\gamma^2}{\nu} \frac{|\tilde{\theta}|^2 |\chi|_{\mathcal{F}}^2}{1 + \nu|\chi|_{\mathcal{F}}^2} \leq \left(\frac{\gamma}{\nu}\right)^2 |\tilde{\theta}|^2 \quad (7.6)$$

which, in view of (7.1), proves (7.2).

iii) By integrating (7.4) over $[0, \infty)$ we obtain

$$\left\| \frac{\epsilon}{\sqrt{1 + \nu|\chi|_{\mathcal{F}}^2}} \right\|_2 \leq \sqrt{V_{\tilde{\theta}}(0)} = \frac{1}{\sqrt{2\gamma}} |\tilde{\theta}(0)|. \quad (7.7)$$

Integration of (7.5) over $[0, \infty)$ and substitution of (7.7) yields (7.3). \square

Remark 7.1: The only difference in the case of the normalized least-squares is that (7.3) becomes $\|\dot{\tilde{\theta}}\|_2 \leq \sqrt{\gamma/\nu} |\tilde{\theta}(0)|$. \square

Theorem 7.1: In the adaptive system (2.1), (3.1) using either the identifier (5.1), (5.2), (5.7) or (5.10), (5.11), (5.16) with normalized update laws, the following inequalities hold

$$i) \quad |z(t)| \leq \frac{|\tilde{\theta}(0)|}{2\sqrt{c_0}} \left(\frac{1}{\kappa_0} + \frac{\gamma^2}{g_0\nu^2} \right)^{1/2} + |z(0)|e^{-c_0 t}, \quad (7.8)$$

$$ii) \quad \left(\frac{1}{t} \int_0^t |z(\tau)|^2 d\tau \right)^{1/2} \leq \frac{|\tilde{\theta}(0)|}{2\sqrt{c_0}} \left(\frac{1}{\kappa_0} + \frac{1}{t} \frac{\gamma^2}{2g_0\nu^2} \right)^{1/2} + \frac{1}{\sqrt{2c_0}} |z(0)|. \quad (7.9)$$

Proof: i) This bound follows by substituting (7.1) and (7.2) into (3.10).

ii) By integrating the first line of (3.13) we get

$$\int_0^t |z(\tau)|^2 d\tau \leq \frac{1}{2c_0} |z(0)|^2 + \frac{1}{4c_0\kappa_0} \|\tilde{\theta}\|_\infty^2 t + \frac{1}{2g_0} \int_0^t \left(\int_0^\tau e^{-2c_0(\tau-s)} |\dot{\hat{\theta}}(s)|^2 ds \right) d\tau. \quad (7.10)$$

Now, to arrive at (7.9), the sequence of integration in (7.10) is interchanged as in the proof of Lemma A.1.(ii).

Remark 7.2: Although the initial states $z_2(0), \dots, z_p(0)$ may depend on c_i, κ_i, g_i , this dependence can be removed by setting $z(0) = 0$ with the following initialization of the reference model

$$x_{m,i}(0) = x_i(0) - \alpha_{i-1}(\bar{x}_{i-1}(0), \hat{\theta}(0), \bar{x}_{m,i-1}(0)). \quad (7.11)$$

It can also be proven that in this initialization $x_m(0)$ does not depend on c_i, κ_i, g_i . Therefore, the bounds (7.8), (7.9) can be made as small as desired by increasing c_0 , and/or κ_0, g_0 . A practical limit to the increase of these gain coefficients is that, in the presence of an error in the initial state measurement, they increase $z(0)$ and the performance deteriorates. As for the pure-feedback systems mentioned in Remark 6.1, the feasibility region may, in general, decrease as c_i, κ_i, g_i increase. \square

Remark 7.3: The above bounds are readily modified to also cover the case when $\tilde{e}(0) \neq 0$. For example, the \mathcal{L}_∞ bound (7.8) is augmented by the term

$$\frac{1}{2\sqrt{c_0}} \left[\frac{\gamma}{4c_0} \left(\frac{1}{\kappa_0} + \frac{2\gamma^2}{g_0\nu^2} \right) + \frac{2\gamma^2}{g_0\nu} \right]^{1/2} |\tilde{e}(0)|. \quad (7.12)$$

If we set both filter initial conditions to zero, namely, $\chi(0) = 0$ and $\chi_0(0) = 0$, we get $\tilde{e}(0) = z(0)$. (This initialization is always exact because it does not depend on the measured state.) In this case, (7.12) shows that, for $z(0) \neq 0$ the performance bound (7.8) is

$$|z(t)| \leq \frac{|\tilde{\theta}(0)|}{2\sqrt{c_0}} \left(\frac{1}{\kappa_0} + \frac{2\gamma^2}{g_0\nu^2} \right)^{1/2} + \frac{1}{2\sqrt{c_0}} \left[\frac{\gamma}{4c_0} \left(\frac{1}{\kappa_0} + \frac{2\gamma^2}{g_0\nu^2} \right) + \frac{2\gamma^2}{g_0\nu} \right]^{1/2} |z(0)| + |z(0)|e^{-c_0 t}. \quad (7.13)$$

\square

Lemma 7.2: For the adaptive system (2.1), (3.1), (5.1), (5.2), (5.7), the following inequalities hold

$$i) \quad \|\chi\|_\infty \leq \frac{1}{2\sqrt{c_0\kappa_0}} \quad (7.14)$$

$$ii) \quad \|\epsilon\|_\infty \leq \frac{|\tilde{\theta}(0)|}{2\sqrt{c_0\kappa_0}} \quad (7.15)$$

$$iii) \quad \|\epsilon\|_2 \leq \sqrt{\frac{1}{2\gamma} + \frac{\nu}{2\gamma} \frac{1}{4c_0\kappa_0}} |\tilde{\theta}(0)|. \quad (7.16)$$

Proof: i) By Lemma A.1-i), and since $\chi(0) = 0$, inequality (6.1) is rewritten as

$$|\chi(t)|_\mathcal{F}^2 \leq |\chi(0)|_\mathcal{F}^2 e^{-2c_0 t} + \int_0^t e^{-2c_0(t-\tau)} \frac{1}{2\kappa_0} d\tau \leq \frac{1}{4c_0\kappa_0} \quad (7.17)$$

and (7.14) follows.

ii) Now (5.5) implies $\|\epsilon\|_\infty \leq \|\chi\|_\infty \|\tilde{\theta}\|_\infty \leq (1/2\sqrt{c_0\kappa_0}) |\tilde{\theta}(0)|$ which proves (7.15).

iii) The bound on the \mathcal{L}_2 norm of ϵ is obtained using

$$\int_0^\infty |\epsilon(\tau)|^2 d\tau \leq \int_0^\infty \frac{|\epsilon(\tau)|^2}{1 + \nu \|\chi\|_\mathcal{F}^2} (1 + \nu \|\chi\|_\mathcal{F}^2) d\tau \leq (1 + \nu \|\chi\|_\mathcal{F}^2) \left\| \frac{\epsilon}{\sqrt{1 + \nu \|\chi\|_\mathcal{F}^2}} \right\|_2. \quad (7.18)$$

By substituting (7.7) and (7.14) into (7.18) we prove (7.16). \square

Remark 7.4: With the bounds (7.14)–(7.16) for the z -swapping scheme we can tighten the bounds on $\|\tilde{\theta}\|_2$ and $\|\tilde{\theta}\|_\infty$ in Lemma 7.1 and make them valid for the unnormalized update laws with $\nu = 0$. It is straightforward to show that $1/\nu$ in (7.2)–(7.3) can be replaced by $\min\{1/\nu, 1/4c_0\kappa_0\}$. The same is true for (7.8)–(7.9). We can also show that for the x -swapping scheme $1/\nu$ can be replaced by $\min\{1/\nu, (p/2\lambda)[\lambda_{\max}(P_0)/\lambda_{\min}(P_0)]\}$. \square

Theorem 7.2 In the adaptive system (2.1), (3.1) with the z -swapping identification scheme (5.1), (5.2), (5.7) the \mathcal{L}_2 norm of z is bounded by

$$\|z\|_2 \leq \frac{|\tilde{\theta}(0)|}{\sqrt{2c_0}} \left(\frac{\gamma}{g_0\nu} + \frac{\gamma}{2c_0^2\kappa_0\nu} + \frac{\nu}{2\kappa_0\gamma} \right)^{1/2} + \frac{|\tilde{\theta}(0)|}{\sqrt{2\gamma}} + \frac{1}{\sqrt{c_0}} |z(0)|. \quad (7.19)$$

Proof: We will calculate the \mathcal{L}_2 norm bound for z as $\|z\|_2 \leq \|\epsilon\|_2 + \|\psi\|_2$, where

$$\psi \triangleq \chi_0 - \chi^T \hat{\theta}. \quad (7.20)$$

A bound on $\|\epsilon\|_2$ is given by (7.16). To obtain a bound on $\|\psi\|_2$, we examine

$$\dot{\psi} = A_z(z, \hat{\theta}, t)\psi - D(z, \hat{\theta}, t)^T \dot{\hat{\theta}} - \chi^T \dot{\hat{\theta}}. \quad (7.21)$$

By using (3.3) and repeating the sequence of inequalities (3.11), we derive

$$\begin{aligned} \frac{d}{dt} \left(\frac{1}{2} |\psi|^2 \right) &\leq -c_0 |\psi|^2 + \frac{1}{4g_0} |\dot{\hat{\theta}}|^2 - \psi^T \chi^T \dot{\hat{\theta}} \\ &\leq -\frac{c_0}{2} |\psi|^2 - \frac{c_0}{2} \left| \psi - \frac{1}{c_0} \chi^T \dot{\hat{\theta}} \right|^2 \\ &\quad + \frac{1}{4g_0} |\dot{\hat{\theta}}|^2 + \frac{1}{2c_0} |\chi^T \dot{\hat{\theta}}|^2 \end{aligned} \quad (7.22)$$

which gives

$$\frac{d}{dt} \left(\frac{1}{2} |\psi|^2 \right) \leq -\frac{c_0}{2} |\psi|^2 + \frac{1}{2} \left(\frac{1}{2g_0} |\dot{\hat{\theta}}|^2 + \frac{1}{c_0} |\chi^T \dot{\hat{\theta}}|^2 \right). \quad (7.23)$$

applying Lemma A.1.(ii) to (7.23), we arrive at

$$\|\psi\|_2 \leq \frac{1}{\sqrt{c_0}} \left(|\psi(0)| + \frac{1}{\sqrt{2g_0}} \|\dot{\hat{\theta}}\|_2 + \frac{1}{\sqrt{c_0}} \|\chi^T \dot{\hat{\theta}}\|_2 \right). \quad (7.24)$$

substituting (7.3) and (7.14) into (7.24) we get

$$\begin{aligned} \|\psi\|_2 &\leq \frac{1}{\sqrt{c_0}} \left(|\psi(0)| + \frac{1}{\sqrt{2g_0}} \|\dot{\hat{\theta}}\|_2 + \frac{1}{\sqrt{c_0}} \|\chi\|_{\mathcal{F}} \|\dot{\hat{\theta}}\|_2 \right) \\ &\leq \frac{|\hat{\theta}(0)|}{2\sqrt{c_0}} \sqrt{\frac{\gamma}{\nu}} \left(\frac{1}{\sqrt{g_0}} + \frac{1}{\sqrt{2c_0^2 \kappa_0}} \right) + \frac{1}{\sqrt{c_0}} |z(0)| \end{aligned} \quad (7.25)$$

where we have assumed that $\chi_0(0) = z(0)$. Combining this and (7.16), and rearranging the terms, we obtain (7.19). \square

The form of the bound (7.19) is favorable because it is linear in $|\hat{\theta}(0)|$. It may not be possible to make the \mathcal{L}_2 norm of z as small as desired by c_0 alone because of the term $|\hat{\theta}(0)|/\sqrt{2\gamma}$. With the standard initialization $z(0) = 0$, however, a possibility to improve the \mathcal{L}_2 performance is by simultaneously increasing c_0 , g_0 and γ .

VIII. RAPPROCHEMENT WITH LINEAR DESIGNS

A connection of the adaptive nonlinear ISS-design presented in this paper with linear estimation-based designs will become clearer when the ISS-controller of Section III is replaced by the weaker SG-controller developed in this section. The only difference between the two controllers is that the SG-controller employs weaker nonlinear damping functions s_i . For example, for an uncertain term $\theta\varphi(x_1)$ in the first equation of the plant, the ISS and SG nonlinear damping functions are respectively

$$\begin{aligned} s_1^{\text{ISS}}(x_1) &= \varphi(x_1)^2 \quad \text{and} \\ s_1^{\text{SG}}(x_1) &= \left[\frac{\varphi(x_1) - \varphi(0)}{x_1} \right]^2 \triangleq \omega(x_1)^2. \end{aligned} \quad (8.1)$$

In this way the growth of s_1^{SG} is reduced by a factor of x_1^2 . In the process of backstepping this reduction is even more pronounced. However, the SG-controller can no longer guarantee the ISS property with respect to $\hat{\theta}$ and $\dot{\hat{\theta}}$. Instead, we reveal a small gain property and prove boundedness with a linear-like Gronwall lemma argument. The main interest in the SG-controller is that for linear systems it becomes linear in x and in that sense is similar to linear estimation-based designs. In contrast, the ISS-controller for linear systems remains nonlinear.

To derive the nonlinear damping expressions for the SG-controller we rewrite the regressor vectors w_i as follows

$$w_i(\bar{z}_i, \hat{\theta}, t) = w_i(0, \hat{\theta}, t) + \omega_i(\bar{z}_i, \hat{\theta}, t)^T \bar{z}_i, \quad (8.2)$$

where $\bar{z}_i \triangleq [z_1, \dots, z_i]^T$, and $\omega_i: \mathbf{R}^i \times \mathbf{R}^p \times \mathbf{R}_+ \rightarrow \mathbf{R}^{i \times p}$ is a matrix-valued function smooth in the first two arguments and continuous and bounded in the third argument (with a slight abuse of notation relative to (3.1) we now express w_i as a function of $\bar{z}_i, \hat{\theta}, t$). Thus we have

$$W = W_0 + [\omega_1^T \bar{z}_1, \dots, \omega_n^T \bar{z}_n] \quad (8.3)$$

where W_0 denotes $W(0, \hat{\theta}, t)$. Likewise, we rewrite (3.3)

$$D = D_0 + [0, \delta_2^T \bar{z}_1, \dots, \delta_n^T \bar{z}_{n-1}] \quad (8.4)$$

where $\delta_i: \mathbf{R}^{i-1} \times \mathbf{R}^p \times \mathbf{R}_+ \rightarrow \mathbf{R}^{(i-1) \times p}$ are matrix-valued functions smooth in the first two arguments and continuous and bounded in the third argument, and D_0 denotes $D(0, \hat{\theta}, t)$.

The SG-controller has the same form (3.1) as the ISS controller, but its nonlinear damping functions are defined as

$$s_i = \kappa_i |\omega_i|_{\mathcal{F}}^2 + g_i |\delta_i|_{\mathcal{F}}^2. \quad (8.5)$$

As in linear estimation-based adaptive control, the SG-controller employs an identifier with normalized update laws.

Theorem 8.1: All the signals in the adaptive system consisting of the plant (2.1), SG-controller (3.1) with (8.5), z -swapping filters (5.1), (5.2), and either the gradient (5.7) or the least-squares (5.8) normalized update law ($\nu > 0$), are globally uniformly bounded for all $t \geq 0$, and $\lim_{t \rightarrow \infty} z(t) = 0$. This means, in particular, that global asymptotic tracking is achieved: $\lim_{t \rightarrow \infty} [y(t) - y_r(t)] = 0$. Furthermore, if $\lim_{t \rightarrow \infty} r(t) = 0$ and $\phi(0) = 0$ then $\lim_{t \rightarrow \infty} x(t) = 0$.

Proof: Using (8.3) we write (5.2) as

$$\dot{\chi}^T = A_z \chi^T + W_0^T + [\omega_1^T \bar{z}_1, \dots, \omega_n^T \bar{z}_n]^T. \quad (8.6)$$

In a fashion similar to (6.1) we compute

$$\begin{aligned} \frac{d}{dt} \left(\frac{1}{2} |\chi|_{\mathcal{F}}^2 \right) &\leq -c_0 |\chi|_{\mathcal{F}}^2 - \sum_{i=1}^n \kappa_i |\omega_i|_{\mathcal{F}}^2 |\chi_i|^2 \\ &\quad + \sum_{i=1}^n \bar{z}_i^T \omega_i \chi_i + \text{tr} \{ W_0^T \chi \} \\ &\leq -\frac{c_0}{2} |\chi|_{\mathcal{F}}^2 + \sum_{i=1}^n \frac{1}{4\kappa_i} |\bar{z}_i|^2 + \frac{1}{2c_0} |W_0|_{\mathcal{F}}^2 \\ &\leq -\frac{c_0}{2} |\chi|_{\mathcal{F}}^2 + \frac{1}{4\kappa_0} |z|^2 + \frac{1}{2c_0} |W_0|_{\mathcal{F}}^2. \end{aligned} \quad (8.7)$$

On the other hand, using (8.4) we write (7.21) as

$$\dot{\psi} = A_z \psi - (\chi + D_0)^T \dot{\hat{\theta}} - [0, \delta_2^T \bar{z}_1, \dots, \delta_n^T \bar{z}_{n-1}]^T \dot{\hat{\theta}} \quad (8.8)$$

and compute

$$\begin{aligned} \frac{d}{dt} \left(\frac{1}{2} |\psi|^2 \right) &\leq -c_0 |\psi|^2 - \sum_{i=1}^n g_i |\delta_i|_{\mathcal{F}}^2 \psi_i^2 \\ &\quad - \psi^T (\chi + D_0)^T \dot{\hat{\theta}} - \sum_{i=1}^n \bar{z}_{i-1}^T \delta_i \psi_i \dot{\hat{\theta}} \\ &\leq -\frac{c_0}{2} |\psi|^2 + \frac{1}{2c_0} |(\chi + D_0)^T \dot{\hat{\theta}}|^2 \\ &\quad + \sum_{i=1}^n \frac{1}{4g_i} |\bar{z}_{i-1}|^2 |\dot{\hat{\theta}}|^2 \\ &\leq -\frac{c_0}{2} |\psi|^2 + \frac{1}{c_0} |\chi|_{\mathcal{F}}^2 |\dot{\hat{\theta}}|^2 \\ &\quad + \frac{1}{4g_0} |z|^2 |\dot{\hat{\theta}}|^2 + \frac{1}{c_0} |D_0|_{\mathcal{F}}^2 |\dot{\hat{\theta}}|^2. \end{aligned} \quad (8.9)$$

The system (8.7), (8.9) is summarized as

$$\frac{d}{dt}(|\chi|_{\mathcal{F}}^2) \leq -c_0|\chi|_{\mathcal{F}}^2 + \frac{1}{2\kappa_0}|z|^2 + \frac{1}{c_0}|W_0|_{\mathcal{F}}^2 \quad (8.10)$$

$$\begin{aligned} \frac{d}{dt}(|\psi|^2) &\leq -c_0|\psi|^2 + \frac{2}{c_0}|\dot{\theta}|^2|\chi|_{\mathcal{F}}^2 + \frac{1}{2g_0}|\dot{\theta}|^2|z|^2 \\ &\quad + \frac{2}{c_0}|D_0|_{\mathcal{F}}^2|\dot{\theta}|^2. \end{aligned} \quad (8.11)$$

From Lemma 5.1 we have $\hat{\theta} \in \mathcal{L}_\infty[0, t_f]$, so $|W_0|_{\mathcal{F}}^2 < k$ and $|D_0|_{\mathcal{F}}^2 < k$, where k denotes a generic positive finite constant. From Lemma 5.1 we also have $\dot{\theta}, \epsilon/\sqrt{1+\nu}|\chi|_{\mathcal{F}}^2 \in \mathcal{L}_2[0, t_f] \cap \mathcal{L}_\infty[0, t_f]$. Let us denote by l_1 a generic function in $\mathcal{L}_1[0, t_f] \cap \mathcal{L}_\infty[0, t_f]$. Since $\epsilon = z + \psi$, then we have

$$\begin{aligned} |z|^2 &\leq 2\frac{|\epsilon|^2}{1+\nu|\chi|_{\mathcal{F}}^2}(1+\nu|\chi|_{\mathcal{F}}^2) + 2|\psi|^2 \\ &\leq 2l_1(1+\nu|\chi|_{\mathcal{F}}^2) + 2|\psi|^2. \end{aligned} \quad (8.12)$$

Thus (8.10)–(8.11) become

$$\frac{d}{dt}(|\chi|_{\mathcal{F}}^2) \leq -(c_0 - l_1)|\chi|_{\mathcal{F}}^2 + \frac{1}{\kappa_0}|\psi|^2 + k \quad (8.13)$$

$$\frac{d}{dt}(|\psi|^2) \leq -(c_0 - l_1)|\psi|^2 + l_1|\chi|_{\mathcal{F}}^2 + k. \quad (8.14)$$

This is a loop with small gain because $|\chi|_{\mathcal{F}}^2$ appears multiplied by l_1 in (8.14). To finish the proof we define the “superstate”

$$X \triangleq |\chi|_{\mathcal{F}}^2 + \frac{2}{\kappa_0 c_0}|\psi|^2 \quad (8.15)$$

differentiate it and substitute (8.13)–(8.14). After straightforward rearrangements and majorizations, we get

$$\dot{X} \leq -\left(\frac{c_0}{2} - l_1\right)X + k. \quad (8.16)$$

By applying the Gronwall lemma we conclude that X is uniformly bounded on $[0, t_f]$. In view of (5.5), ϵ is bounded, which along with the boundedness of ψ , proves that z is bounded. Thus $t_f = \infty$. The rest of the proof is the same as for Theorem 6.1, and again uses Lemma 4.2 for proving convergence. \square

In contrast to the ISS design, the global result has been established only with normalized update laws. The issue of normalization in adaptive nonlinear design was discussed in [26].

Performance bounds similar to those in Section VII for the ISS design are not available for the SG design, nor is it clear how to develop its x -swapping version.

In Section IX we compare the linear SG design with the new ISS design for a linear plant.

IX. EXAMPLES AND DISCUSSION

The first example in this section illustrates the performance properties of the ISS design on the relative-degree two plant

$$\begin{aligned} \dot{x}_1 &= x_2 + \theta\varphi(x_1) \\ \dot{x}_2 &= u \end{aligned} \quad (9.1)$$

with the objective to regulate x to zero (without a reference model). The second example makes a comparison between the ISS design and the SG design.

We define the error variables

$$\begin{aligned} z_1 &= x_1 \\ z_2 &= x_2 - \alpha_1(x_1, \hat{\theta}). \end{aligned} \quad (9.2)$$

The two-step ISS-controller design

$$\begin{aligned} \alpha_1 &= -c_1 z_1 - \kappa_1 \varphi^2 z_1 - \hat{\theta} \varphi \\ u &= -z_1 - c_2 z_2 - \kappa_2 \left(\frac{\partial \alpha_1}{\partial x_1} \right)^2 \varphi^2 z_2 - g_2 \left(\frac{\partial \alpha_1}{\partial \hat{\theta}} \right)^2 z_2 \\ &\quad + \frac{\partial \alpha_1}{\partial x_1} (x_2 + \hat{\theta} \varphi) \end{aligned} \quad (9.3)$$

results in the error system

$$\dot{z} = A_z(z, \hat{\theta})z + \begin{bmatrix} -\frac{\varphi}{\partial x_1} \\ 0 \end{bmatrix} \dot{\hat{\theta}} + \begin{bmatrix} 0 \\ -\frac{\partial \alpha_1}{\partial \hat{\theta}} \end{bmatrix} \dot{\hat{\theta}} \quad (9.4)$$

where

$$A_z = \begin{bmatrix} -c_1 - \kappa_1 \varphi^2 & \\ -1 & -c_2 - \kappa_2 \left(\frac{\partial \alpha_1}{\partial x_1} \right)^2 \varphi^2 - g_2 \left(\frac{\partial \alpha_1}{\partial \hat{\theta}} \right)^2 \end{bmatrix} \quad (9.5)$$

The z -swapping identifier is designed with the following filters

$$\chi_0 = A_z(z, \hat{\theta})\chi_0 + \begin{bmatrix} \frac{\varphi}{\partial x_1} \\ 0 \end{bmatrix} \hat{\theta} - \begin{bmatrix} 0 \\ -\frac{\partial \alpha_1}{\partial \hat{\theta}} \end{bmatrix} \hat{\theta} \quad (9.6)$$

$$\chi^T = A_z(z, \hat{\theta})\chi^T + \begin{bmatrix} \frac{\varphi}{\partial x_1} \\ 0 \end{bmatrix} \hat{\theta} \quad (9.7)$$

which are used to implement the augmented error

$$\epsilon = z + \chi_0 - \chi^T \hat{\theta} \quad (9.8)$$

and the gradient update law

$$\dot{\hat{\theta}} = \gamma \frac{\chi \epsilon}{1 + \nu \chi \chi^T}. \quad (9.9)$$

The x -swapping identifier is designed with the following filters

$$\Omega_0 = -(\bar{\alpha} + \lambda \varphi^2)(\Omega_0 - x_1) + x_2, \quad \Omega_0 \in \mathbb{R} \quad (9.10)$$

$$\dot{\Omega} = -(\bar{\alpha} + \lambda \varphi^2)\Omega + \varphi, \quad \Omega \in \mathbb{R} \quad (9.11)$$

which are used to implement the equation error

$$\epsilon = x_1 - \Omega_0 - \Omega \theta \quad (9.12)$$

and the gradient update law

$$\dot{\hat{\theta}} = \gamma \frac{\Omega \epsilon}{1 + \nu \Omega^2}. \quad (9.13)$$

This reveals that the x -swapping approach is uncertainty specific in the sense that only the terms φ , multiplying the unknown parameter θ need to be filtered. This opens a possibility for a reduction in the dynamic order of the identifier.

In simulations, the only difference between the z -swapping and the x -swapping approach was in the value of γ needed to achieve the same speed of adaptation—higher value was needed in the z -swapping case. Since the responses were similar we show them only for the z -swapping scheme.

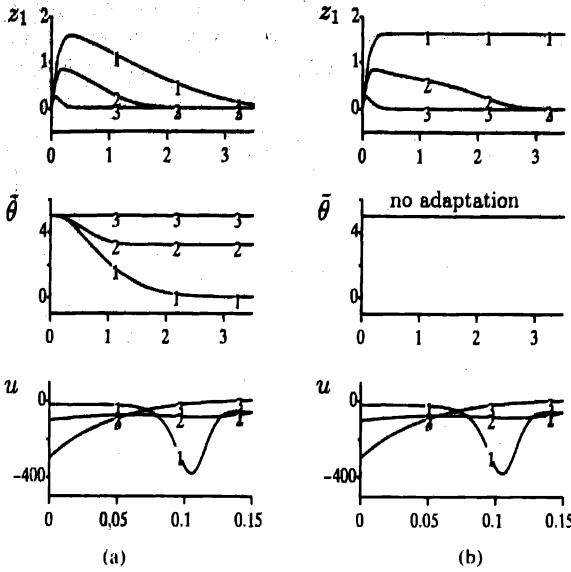


Fig. 1. Dependence of the transients on c_0 with $\kappa_0 = g_0 = 1$. (Note an expanded time scale for control u .) (a) $\gamma = 10$; (b) $\gamma = 0$. 1: $c_0 = 1$; 2: $c_0 = 5$; 3: $c_0 = 15$.

Example 8.1 (ISS-Performance): We consider system (9.1) with nonlinearity $\varphi(x_1) = x_1^2$. The simulations were carried out with nominal values $c_1 = c_2 = c_0 = \kappa_1 = \kappa_2 = \kappa_0 = g_2 = g_0 = 1$, $\gamma = 10$, $\theta = 5$, $\theta(0) = 0$ which were judged to give representative responses. All simulations are with following initial conditions: $x(0) = -\chi_0$ ($0 = [0, 10]^T$), $\chi(0) = 0$ (to set $\tilde{c}(0) = 0$).

Fig. 1(a) illustrates Theorems 7.1 and 7.2. The design parameter c_0 can be used for systematically improving the transient performance. Up to a certain point the error transients and the control effort in Fig. 1(a) are simultaneously decreasing as c_0 increases. Beyond that point the control effort starts increasing. The control u is given in an expanded time scale in order to clearly display the main qualitative differences among the three cases. Fig. 1(b) illustrates Corollary 3.1. When adaptation is switched off, the states are uniformly bounded and converge to (or remain inside) a compact residual set. Corollary 3.1 does not describe the behavior inside the residual set, which may contain multiple equilibria, limit cycles, etc. For this example (but not in general), there is an asymptotically stable equilibrium at the origin for any value of the parameter error. For small values of c_0 , this equilibrium has a basin of attraction which is strictly inside the residual set. For higher values of c_0 the global asymptotic stability is achieved.

Fig. 2 shows the influence of κ_0 on transients. According to Theorem 7.1 and Remark 7.4, the peak values can be decreased by increasing κ_0 , which is confirmed by the plot. The L_2 performance may not be improved, however, by increasing κ_0 because the κ -terms slow down the adaptation and make the transients longer. The effect of the g -terms was shown to be significant only for very small c_0 and κ_0 or for very large γ .

Fig. 3 demonstrates the influence of the adaptation gain γ on transients. Due to the slow initial adaptation, which should be attributed not only to the normalized gradient update law but also to the fact that the regressor is filtered, there is a clear separation of action of the nonadaptive controller, which

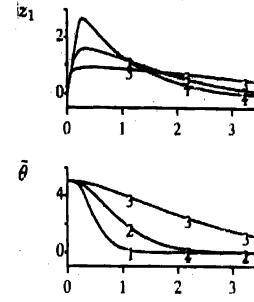


Fig. 2. Dependence of the transients on κ_0 with $c_0 = g_0 = 1$, $\gamma = 10$. 1: $c_0 = 1$; 2: $c_0 = 5$; 3: $c_0 = 15$.

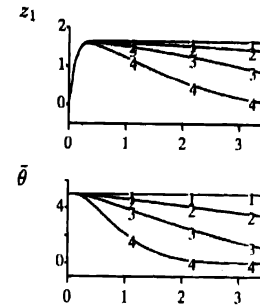


Fig. 3. Dependence of the transients on γ with $c_0 = g_0 = \kappa_0 = 1$. 1: $c_0 = 1$; 2: $c_0 = 5$; 3: $c_0 = 15$.

at the beginning brings the state z quickly to the residual set, and the adaptive controller which takes over to drive the state to the origin. The property that the L_∞ bounds are increasing functions of γ , to be expected from Theorem 7.1, was exhibited in simulations only with extremely high values of γ . This indicates that some of the bounds derived are not very tight over the entire range of design parameter values.

Finally, an explanation is in order about the initial condition $x(0)$ in our simulations. We used $x(0) = [0, 10]^T$, and hence $z(0) = [0, 10]^T$ which is independent of the design gains c_0, κ_0, g_0 . This is why the peak of z_1 decreases monotonically as any of these gains increases. If, instead, we used $x_1(0) \neq 0$, then, according to Remark 7.2, we would have added an appropriately initialized reference model (with $r(t) \equiv 0$). In this way, bad transients would be eliminated by following a less aggressive path to the origin. \square

Example 8.2 (ISS vs. SG): Let us consider system (9.1) with $\varphi(x_1) = x_1$. For this linear system we make a comparison between the ISS and SG designs. The only difference is that the terms $\kappa_1 \varphi^2, \kappa_2 (\partial \alpha_1 / \partial x_1)^2 \varphi^2, g_2 (\partial \alpha_1 / \partial \theta)^2$ in the ISS design are, respectively, replaced by $\kappa_1, \kappa_2 (\partial \alpha_1 / \partial x_1)^2, g_2$ in the SG design. The same design coefficients and initial conditions are used as in Example 8.1, except for $\theta = 3$. The adaptation gains, $\gamma = 5$ for the ISS design, and $\gamma = 1.5$ for the SG design, are chosen so that the rate of parameter convergence is the same for both designs. The control law of the SG design is linear in x and nonlinear in $\hat{\theta}$.

Fig. 4 shows the difference in performance between the two designs. The ISS design uses larger control effort and achieves better attenuation of the z_1 -transient. The dashed responses illustrate the underlying nonadaptive behavior ($\gamma = 0$). While

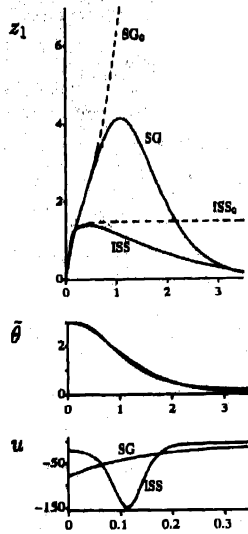


Fig. 4. ISS design vs. SG design. The dashed lines show $z_1(t)$ when $\gamma = 0$. (Note an expanded time scale for control u .)

the response of the SG design exhibits linear exponential instability, the response of the ISS design, according to Corollary 3.1, is bounded. Hence, there is a clear trade-off of performance improvement versus control effort between the new ISS design and the linear adaptive designs such as the SG design. \square

X. CONCLUSIONS

Recent Lyapunov-based recursive designs of adaptive controllers for nonlinear systems transformable into the parametric-strict-feedback form [12], [8], [19], [36] achieve global stability and tracking, but do not allow a choice of parameter update laws. In these designs the wealth of knowledge about standard identifiers is not utilized because the identifier does not appear as a separate module of the adaptive system.

A complete separation of the controller and identifier modules is one of the main accomplishments of this paper. It has been achieved by a new nonlinear controller with an input-to-state stability property with respect to the parameter estimation error and its derivative as disturbance inputs. This strong ISS-controller remains nonlinear even when the plant is linear. For comparison with linear estimation-based designs, a weaker SG-controller is introduced, resulting in a small-gain rather than the ISS property. For linear plants this controller is linear.

As a separate module, the ISS-controller can be connected with the standard unnormalized or normalized gradient or least-squares identifiers, while the SG-controller requires normalization. The connection of the controller and identifier modules is made possible by a nonlinear extension of the well known swapping lemma.

In addition to the global boundedness and tracking, the new design also provides explicit bounds on the transient performance, which can be utilized for its systematic improvement.

The results of this paper assume that the full state is available for feedback. Relying on the experience gained with recent recursive output-feedback designs, such as [22], [14], it

is expected that the estimation-based design of this paper will be extended to nonlinear systems in the output-feedback form.

The applicability of various designs, Lyapunov-based or estimation-based, will ultimately depend on their robustness with respect to unmodeled phenomena. This is another important topic of current research.

APPENDIX

Lemma A.1: Let $v, \rho: \mathbb{R}_+ \rightarrow \mathbb{R}$, $c, b > 0$. If

$$\dot{v} \leq -cv + b\rho^2, \quad v(0) \geq 0 \quad (\text{A.1})$$

i) then

$$v(t) \leq v(0)e^{-ct} + b \int_0^t e^{-c(t-\tau)} \rho(\tau)^2 d\tau. \quad (\text{A.2})$$

ii) If, in addition, $\rho \in \mathcal{L}_2$, then $v \in \mathcal{L}_\infty \cap \mathcal{L}_1$ and

$$\|v\|_1 \leq \frac{1}{c}(v(0) + b\|\rho\|_2^2). \quad (\text{A.3})$$

Proof: i) Upon multiplication of (A.1) by e^{ct} , it becomes

$$\frac{d}{dt}(v(t)e^{ct}) \leq b\rho(t)^2 e^{ct}. \quad (\text{A.4})$$

Integrating (A.4) over $[0, t]$, we arrive at (A.2).

ii) Noting that (A.2) implies that

$$v(t) \leq v(0)e^{-ct} + b \sup_{\tau \in [0, t]} \{e^{-c(t-\tau)}\} \int_0^t \rho(\tau)^2 d\tau \quad (\text{A.5})$$

we conclude that $v \in \mathcal{L}_\infty$. By integrating (A.2) over $[0, t]$, we get

$$\begin{aligned} \int_0^t v(\tau) d\tau &\leq \int_0^t v(0)e^{-c\tau} d\tau + b \int_0^t \left[\int_0^\tau e^{-c(\tau-s)} \rho(s)^2 ds \right] d\tau \\ &\leq \frac{1}{c}v(0) + b \int_0^t e^{cs} \rho(s)^2 \left(\int_s^t e^{-c\tau} d\tau \right) ds \\ &\leq \frac{1}{c}v(0) + b \int_0^t e^{cs} \rho(s)^2 \frac{1}{c} e^{-cs} ds \\ &\leq \frac{1}{c} \left[v(0) + b \int_0^t \rho(\tau)^2 d\tau \right] \end{aligned} \quad (\text{A.6})$$

which proves (A.3). \square

Lemma A.2: Let $T_i: u \mapsto \zeta_i$, $i = 1, 2$ be linear time-varying operators defined by

$$\dot{\zeta}_i = A_i(t)\zeta_i + u \quad (\text{A.7})$$

where $A_i: \mathbb{R}_+ \rightarrow \mathbb{R}^{n \times n}$ are continuous, bounded and exponentially stable. Suppose $\tilde{\theta}: \mathbb{R}_+ \rightarrow \mathbb{R}^p$ is differentiable, $\phi: \mathbb{R}_+ \rightarrow \mathbb{R}^{p \times m}$ is piecewise continuous and bounded, and $M: \mathbb{R}_+ \rightarrow \mathbb{R}^{n \times n}$ is bounded and has a bounded derivative on \mathbb{R}_+ . If $\tilde{\theta} \in \mathcal{L}_2$ then

$$T_1[\phi^T]\tilde{\theta} \in \mathcal{L}_2 \Rightarrow T_2[M\phi^T]\tilde{\theta} \in \mathcal{L}_2. \quad (\text{A.8})$$

If moreover, $M(t)$ is nonsingular $\forall t$, and M^{-1} is bounded and has a bounded derivative on \mathbb{R}_+ then (A.8) holds in both directions.

Proof: Suppose that $T_1[\phi^T]\tilde{\theta} \in \mathcal{L}_2$. By Lemma 4.2, $T_1[\phi^T]\tilde{\theta} - T_1[\phi^T]\tilde{\theta} \in \mathcal{L}_2$ and therefore $\zeta_1 \triangleq T_1[\phi^T]\tilde{\theta} \in \mathcal{L}_2$. We will show first that $\zeta_2 \triangleq T_2[M\phi^T\tilde{\theta}] \in \mathcal{L}_2$. By substituting $\tilde{\theta} = \zeta_1 - A_1(t)\zeta_1$ into the variation of constants formula and applying partial integration we calculate

$$\begin{aligned} \zeta_2(t) &= \Phi_2(t, 0)\zeta_2(0) + \int_0^t \Phi_2(t, \tau)M(\tau)\phi^T(\tau)\tilde{\theta}(\tau) d\tau \\ &= \Phi_2(t, 0)\zeta_2(0) + \int_0^t \Phi_2(t, \tau)M(\tau) \\ &\quad \cdot [\dot{\zeta}_1(\tau) - A_1(\tau)\zeta_1(\tau)] d\tau \\ &= \Phi_2(t, 0)\zeta_2(0) + M(t)\zeta_1(t) - \Phi_2(t, 0)M(0)\zeta_1(0) \\ &\quad + \int_0^t \Phi_2(t, \tau)[\dot{M}(\tau) + A_2(\tau)M(\tau) - M(\tau)A_1(\tau)] \\ &\quad \cdot \zeta_1(\tau) d\tau \end{aligned} \quad (\text{A.9})$$

where $\Phi_2(t, \tau)$ is the state transition matrix of $A_2(t)$ that satisfies $\|\Phi_2(t, \tau)\|_2 \leq ke^{-\alpha(t-\tau)}$, $k, \alpha > 0$. It is clear that $\Phi_2(t, 0)\zeta_2(0) + M(t)\zeta_1(t) - \Phi_2(t, 0)M(0)\zeta_1(0) \in \mathcal{L}_2$ because $\Phi_2(t, 0)$ is exponentially decaying, $M(t)$ is bounded and $\zeta_1 \in \mathcal{L}_2$. Since

$$\begin{aligned} &\Phi_2(t, \tau)[\dot{M}(\tau) + A_2(\tau)M(\tau) - M(\tau)A_1(\tau)]\zeta_1(\tau) d\tau \\ &\leq \|\dot{M} + A_2M - MA_1\|_\infty^2 k^2 \int_0^t e^{-2\alpha(t-\tau)} |\zeta_1(\tau)|^2 d\tau \end{aligned} \quad (\text{A.10})$$

then similarly to (4.16)–(4.17) from the proof of Lemma 4.2, we can show that the expression (A.10) is in \mathcal{L}_2 . Thus $\zeta_2 = T_2[M\phi^T\tilde{\theta}] \in \mathcal{L}_2$. By Lemma 4.2, $T_2[M\phi^T\tilde{\theta}] - T_2[M\phi^T]\tilde{\theta} \in \mathcal{L}_2$ and therefore $T_2[M\phi^T\tilde{\theta}] \in \mathcal{L}_2$. The proof of the other direction of (A.8) when $M(t)$ is nonsingular $\forall t$, and M^{-1} is bounded and has a bounded derivative on \mathbb{R}_+ is identical. \square

ACKNOWLEDGMENT

The authors are thankful to L. Praly for his sharp and constructive criticism.

REFERENCES

- [1] A. M. Annaswamy, D. Seto, and J. Baillieul, "Adaptive control of a class of nonlinear systems," in *Proc. 7th Yale Workshop Adaptive and Learning Sys.*, New Haven, CT, 1992.
- [2] G. Bastin and G. Campion, "Indirect adaptive control of linearly parametrized nonlinear systems," in *Proc. 3rd IFAC Sym. Adaptive Sys. Contr. and Sig. Process.*, Glasgow, UK, 1989.
- [3] G. Campion and G. Bastin, "Indirect adaptive state-feedback control of linearly parametrized nonlinear systems," *Int. J. Adaptive Contr. Sig. Process.*, vol. 4, pp. 345–358, 1990.
- [4] A. Datta and P. Ioannou, "Performance analysis and improvement in model reference adaptive control," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 2370–2387, 1994.
- [5] B. Egardt, *Stability of Adaptive Controllers*. New York: Springer Verlag, 1979.
- [6] G. C. Goodwin and D. Q. Mayne, "A parameter estimation perspective of continuous-time model reference adaptive control," *Automatica*, vol. 23, pp. 57–70, 1987.
- [7] P. A. Ioannou and J. Sun, *Stable and Robust Adaptive Control*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
- [8] Z. P. Jiang and L. Praly, "Iterative designs of adaptive controllers for systems with non-linear integrators," in *Proc. 30th IEEE Conf. Decis. Contr.*, Brighton, UK, Dec. 1991, pp. 2482–2487.
- [9] I. Kanellakopoulos, P. V. Kokotović, and R. H. Middleton, "Observer-based adaptive control of nonlinear systems under matching conditions," in *Proc. 1990 American Contr. Conf.*, San Diego, CA, pp. 549–553, 1990.
- [10] ———, "Indirect adaptive output-feedback control of a class of nonlinear systems," in *Proc. 29th IEEE Conf. Decis. Contr.*, Honolulu, HI, Dec. 1990, pp. 2714–2719.
- [11] I. Kanellakopoulos, P. V. Kokotović, and R. Marino, "An extended direct scheme for robust adaptive nonlinear control," *Automatica*, vol. 27, pp. 247–255, 1991.
- [12] I. Kanellakopoulos, P. V. Kokotović, and A. S. Morse, "Systematic design of adaptive controllers for feedback linearizable systems," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1241–1243, 1991.
- [13] ———, "Adaptive output-feedback control of systems with output nonlinearities," *Foundation Adaptive Control*, P. V. Kokotović, Ed., Berlin: Springer-Verlag, 1991, pp. 495–525.
- [14] ———, "Adaptive output-feedback control of a class of nonlinear systems," in *Proc. 30th IEEE Conf. Decis. Contr.*, Brighton, UK, Dec. 1991, pp. 1082–1087.
- [15] ———, "A toolkit for nonlinear feedback design," *Sys. Contr. Lett.*, vol. 18, pp. 83–92, 1992.
- [16] I. Kanellakopoulos, "Passive adaptive control of nonlinear systems," *Int. J. Adaptive Contr. Sig. Process.*, vol. 7, pp. 339–352, 1993.
- [17] P. V. Kokotović, Ed., *Foundation Adaptive Control*. Berlin: Springer-Verlag, 1991.
- [18] ———, "The joy of feedback: Nonlinear and adaptive," *Contr. Sys. Mag.*, vol. 12, pp. 7–17, 1991.
- [19] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, "Adaptive nonlinear control without overparametrization," *Sys. Contr. Lett.*, vol. 19, pp. 177–185, 1992.
- [20] M. Krstić, P. V. Kokotović and I. Kanellakopoulos, "Transient performance improvement with a new class of adaptive controllers," *Sys. Contr. Lett.*, vol. 21, No. 6, 1993.
- [21] R. Marino and P. Tomei, "Global adaptive observers and output-feedback stabilization for a class of nonlinear systems," in *Foundations of Adaptive Control*, P. V. Kokotović, Ed., Berlin: Springer-Verlag, 1991, pp. 455–493.
- [22] ———, "Global adaptive output-feedback control of nonlinear systems, Part I: Linear parametrization," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 17–32, 1993.
- [23] ———, "Global adaptive output-feedback control of nonlinear systems, Part II: Nonlinear parametrization," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 33–49, 1993.
- [24] R. H. Middleton and G. C. Goodwin, "Adaptive computed torque control for rigid link manipulators," *Sys. Contr. Lett.*, vol. 10, pp. 9–16, 1988.
- [25] A. S. Morse, "Global stability of parameter-adaptive control systems," *IEEE Trans. Automat. Contr.*, vol. 25, pp. 433–439, 1980.
- [26] ———, "A comparative study of normalized and unnormalized tuning errors in parameter adaptive control," *Int. J. Adaptive Contr. Sig. Process.*, vol. 6, pp. 309–318, 1992.
- [27] K. Nam and A. Arapostathis, "A model-reference adaptive control scheme for pure-feedback nonlinear systems," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 803–811, 1988.
- [28] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [29] J. B. Pomet and L. Praly, "Indirect adaptive nonlinear control," in *Proc. 27th IEEE Conf. Decis. Contr.*, Austin, TX, Dec. 1988, pp. 2414–2415.
- [30] ———, "Adaptive nonlinear control: an estimation-based algorithm," in *New Trends Nonlinear Contr. Theory*, J. Descusse, M. Fliess, A. Isidori, and D. Leborgne, Eds. New York: Springer-Verlag, 1989.
- [31] ———, "Adaptive nonlinear regulation: estimation from the Lyapunov equation," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 729–740, 1992.
- [32] L. Praly, G. Bastin, J.-B. Pomet, and Z. P. Jiang, "Adaptive stabilization of nonlinear systems," in *Foundations of Adaptive Control*, P. V. Kokotović, Ed., Berlin: Springer-Verlag, 1991, pp. 347–434.
- [33] L. Praly, "Adaptive regulation: Lyapunov design with a growth condition," *Int. J. Adaptive Contr. Sig. Process.*, vol. 6, pp. 329–351, 1992.
- [34] S. S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence and Robustness*, Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [35] S. S. Sastry and A. Isidori, "Adaptive control of linearizable systems," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 1123–1131, 1989.
- [36] D. Seto, A. M. Annaswamy, and J. Baillieul, "Adaptive control of a class of nonlinear systems with a triangular structure," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 1411–1428, 1994.
- [37] E. D. Sontag, "Smooth stabilization implies coprime factorization," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 435–443, 1989.
- [38] D. Taylor, P. V. Kokotović, R. Marino, and I. Kanellakopoulos, "Adaptive regulation of nonlinear systems with unmodeled dynamics," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 405–412, 1989.

- [39] A. R. Teel, R. R. Kadiyala, P. V. Kokotović, and S. S. Sastry, "Indirect techniques for adaptive input-output linearization of non-linear systems," *Int J Contr*, vol. 53, pp. 193-222, 1991.
- [40] A. R. Teel, "Error-based adaptive non-linear control and regions of feasibility," *Int J Adaptive Contr Sig Process*, vol. 6, pp. 319-327, 1992.



Miroslav Krstić (S'92) was born in Pirot, Yugoslavia in 1964. He received the BSEE degree (1989) from the University of Belgrade, Yugoslavia, and the MSEE degree (1992) from the University of California, Santa Barbara, where he has been since 1991, working towards the Ph.D. degree.

He received the 1993 IEEE CDC Best Student Paper Award for the paper "Adaptive nonlinear control with nonlinear swapping" co-authored with P. V. Kokotović.

In 1993 he was also a recipient of the UCSB Engineering Fellowship. His research interests include adaptive control of nonlinear and linear systems, robust nonlinear control, dynamical systems, and control applications.



Petar V. Kokotović (SM'74-F'80) has been active for more than 30 years as a control engineer, researcher, and educator, first in his native Yugoslavia and then, from 1966 through 1990 at the University of Illinois, where he held the endowed Granger Chair. Since 1991 he has been Co-Director (with A. J. Laub) of the newly formed Center for Control Engineering and Computation at the University of California, Santa Barbara. He has coauthored eight books and numerous articles contributing to sensitivity analysis, singular perturbation methods, and robust adaptive and nonlinear control. He is also active in industrial applications of control theory. As a consultant to Ford, he was involved in the development of the first series of automotive computer controls, and at General Electric, he participated in large-scale systems studies.

Dr. Kokotović received the 1990 Quazza Medal, the 1983 and 1993 Outstanding IEEE TRANSACTIONS Paper Awards, and presented the 1991 Bodé Prize Lecture. He is the recipient of the 1995 IEEE Control Systems Award.

Approximate Decoupling and Asymptotic Tracking for MIMO Systems

D. N. Godbole, *Student Member, IEEE*, and S. S. Sastry, *Senior Member, IEEE*

Abstract—This paper presents an algorithm for approximate input-output decoupling of nonlinear MIMO systems that are either numerically ill-posed or exhibit nearly singular behavior in the application of decoupling algorithms. Although the systems considered are regular, so that the exact decoupling algorithms are applicable in this case, they require inversion of an ill-conditioned matrix, and yield high gain feedback solutions that may result in actuator saturation and cancellation of high frequency zeros. The approximate algorithms of this paper are numerically robust, and provide solutions that do not cancel far off right-half plane zeros. This latter characteristic is especially valuable when some of the far off right-half plane zeros are unstable. The algorithms are inspired by and are generalizations of some examples in the flight control literature [1], [2], [3].

I. INTRODUCTION

THE nonlinear control toolbox has grown enormously in the last decade. Central to this development is the theory of feedback linearization for nonlinear systems (see [4], [5]) and the solution to the multi-input multi-output (MIMO) decoupling problem. Several algorithms have been proposed in the literature for solving the problem of exact decoupling for nonlinear MIMO systems, see for example [4]–[9]. These algorithms all require the determination of the inverse of a so-called decoupling matrix. Practical implementation of these algorithms, however, is difficult when the decoupling matrix is ill conditioned or close to singularity, in which case the decoupling of such systems needs excessively large control effort. Further, the algorithm is not numerically robust since it requires the inverse of an ill conditioned matrix.

In this paper, we propose a numerically robust input-output decoupling algorithm for invertible nonlinear MIMO systems. Our efforts are motivated, in part, by the use in Hauser *et al.* [2], the work of Singh [1], [3] of such techniques to aircraft flight control problems and the work of Barbot *et al.* [10] with applications to models of electric motors and the Pelousov–Zhabotinsky reaction kinetics. In these examples, the intuition for approximation and the choice of parameters to be approximated is provided by the physics of the problem. This paper attempts to formalize the theory involved in

these examples and provide an algorithm for more general MIMO nonlinear systems, whose physical derivation may not be explicitly known by the designer. Another recent paper by Grizzle and Di Benedetto [11] provides an approximate decoupling algorithm for systems that are not decouplable by the exact decoupling algorithms by reason of their not being regular. The approximate algorithm in this paper aims at those systems that are regular and so decouplable by the exact decoupling algorithms, but the numerics of the decoupling are poorly conditioned.

In addition to the numerical robustness of our approximate algorithm, it also serves another important purpose: The exact input-output decoupling algorithm is essentially a pole-zero cancelling control law. Thus, this law cancels zeros of the open-loop system regardless of whether they are in the left half plane or the right-half plane and regardless of their magnitude. In particular the input-output decoupling control law can only be applied to minimum phase nonlinear systems. For systems with far off zeros, cancellation may not be necessary, since the zeros do not play a large role in the system dynamics and indeed may cause instability when the zeros lie in the right-half plane. In either case, cancellation of far off zeros results in high gain controllers. Our approximate decoupling algorithm does not cancel the far off zeros of the open-loop system, thereby providing reasonable gain, practically implementable solutions. The price to be paid is the replacement of asymptotically exact tracking control laws by approximate tracking control laws. This connection between regular perturbations of nonlinear systems and the far off zeros was first pointed out in [12] for the single-input single-output (SISO) case and in [13] for the MIMO case. Systems in which these far off zeros are in the right-half plane are called slightly nonminimum phase systems in [2]. The approximate decoupling controller, in this case, results in a stable closed-loop system.

It is possible to develop several different approximate decoupling algorithms starting from the different decoupling algorithms in the literature. The algorithm presented here is based on the Descusse and Moog dynamic decoupling algorithm (see [6], [14]). Section II reviews the Descusse–Moog algorithm. In Section III, we state the approximate algorithm and Section IV compares its convergence properties with that of the Descusse–Moog algorithm. In many cases, input-output decoupling is a first step in designing a tracking controller. Section V examines the effect of the approximate decoupling on the performance of tracking controller.

Manuscript received March 26, 1993. Recommended by Associate Editor, M. Bloch. This research was supported in part by Army Research Office under Grant DAAL-91-G-091 and NASA under Grant NAG 2-243.

The authors are with the Department of Electrical Engineering and Computer Sciences Intelligent Machines and Robotics Laboratory, University of California, Berkeley, CA 94720 USA.
IEEE Log Number 9408267.

II. DECOUPLING ALGORITHMS FOR NONLINEAR SYSTEMS

Consider the square (i.e., number of inputs is equal to the number of outputs) MIMO nonlinear control system described by

$$\begin{aligned}\dot{x} &= f(x) + \sum_{i=1}^m g_i(x) u_i \\ y_j &= h_j(x) \quad j = 1, \dots, m\end{aligned}\quad (1)$$

where $x \in \mathcal{R}^n$, $f(x)$, $g_1(x), \dots, g_m(x)$ are analytic vector fields on \mathcal{R}^n and $h_1(x), \dots, h_m(x)$ analytic functions on \mathcal{R}^n . For convenience, these equations are written as

$$\sum_0: \begin{cases} \dot{x} = f(x) + g(x)u \\ y = h(x). \end{cases}$$

Throughout the analysis, we will assume that x_0 is an equilibrium point of the autonomous system, that is $f(x_0) = 0$. We will assume (without loss of generality) that $h(x_0) = 0$. All the analysis in this paper will be local and will be valid in a given open neighborhood U of x_0 . We now review some algorithms for decoupling of MIMO nonlinear systems.

We assume in what follows that each output y_j has a well defined relative degree γ_j , i.e., there exists an integer γ_j such that

$$L_{g_i} L_f^l h_j(x) \equiv 0 \quad \forall l < \gamma_j - 1, \forall 1 \leq i \leq m, \forall x \in U.$$

Collecting these calculation, we have

$$\begin{aligned} \begin{bmatrix} y_1^{\gamma_1} \\ y_2^{\gamma_2} \\ \vdots \\ y_m^{\gamma_m} \end{bmatrix} &= \begin{bmatrix} L_f^{\gamma_1} h_1(x) \\ L_f^{\gamma_2} h_2(x) \\ \vdots \\ L_f^{\gamma_m} h_m(x) \end{bmatrix} \\ &+ \begin{bmatrix} L_{g_1} L_f^{\gamma_1-1} h_1(x) & \dots & L_{g_m} L_f^{\gamma_1-1} h_1(x) \\ L_{g_1} L_f^{\gamma_2-1} h_2(x) & \dots & L_{g_m} L_f^{\gamma_2-1} h_2(x) \\ \vdots & \dots & \vdots \\ L_{g_1} L_f^{\gamma_m-1} h_m(x) & \dots & L_{g_m} L_f^{\gamma_m-1} h_m(x) \end{bmatrix} u \\ &:= b(x) + A(x)u. \end{aligned}\quad (2)$$

$A(x)$ is called the decoupling matrix. If $A(x)$ is invertible at every point in U , then the static-state feedback given by

$$u = (A(x))^{-1}[-b(x) + v] \quad (3)$$

will result in a closed-loop system that is decoupled from input v to output y . This decoupled and input-output linearized system is given by

$$\begin{bmatrix} y_1^{\gamma_1} \\ y_2^{\gamma_2} \\ \vdots \\ y_m^{\gamma_m} \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}. \quad (4)$$

If the matrix $A(x)$ is singular, we cannot use a static state feedback to decouple the nonlinear system ([4]), and we have

to look for a dynamic state feedback to achieve input-output decoupling.

A. Dynamic Decoupling

If decoupling of the system \sum_0 of (1) cannot be achieved by static state feedback, it may still be possible to find a dynamic compensator of the form

$$\sum_c: \begin{cases} \dot{z} = D(x, z) + E(x, z)v \\ u = F(x, z) + G(x, z)v \end{cases} \quad (5)$$

with $z \in \mathcal{R}^{n_c}$, $v \in \mathcal{R}^m$, such that the closed-loop system denoted by \sum_c (for extended system)

$$\sum_c: \begin{cases} \dot{x} = f(x) + g(x)F(x, z) + g(x)G(x, z)v \\ \dot{z} = D(x, z) + E(x, z)v \\ y = h(x) \end{cases} \quad (6)$$

is decoupled from v to y . The dynamic feedback that decouples the system \sum_0 of (1) is actually a static feedback to decouple the extended system \sum_c of (6). There are a number of algorithms in the literature for dynamic decoupling. The approximate decoupling algorithm we will propose is based on the Descusse and Moog algorithm of ([6], [14]). We will review the original algorithm to fix notation.

Descusse and Moog Dynamic Decoupling Algorithm: Define the extended system at the end of iteration $k-1$ to be \sum_k having x^c as its state and equilibrium point x_0^c . The algorithm will start at $k=0$ with the given system \sum_0 having state $x \in \mathcal{R}^n$ and x_0 as its equilibrium point. The outputs of the system are unchanged during the course of the algorithm.

Step i: Compute the relative degrees $\gamma_i^k (i \in \{1, \dots, m\})$ for the m outputs of \sum_k . Define the decoupling matrix $A_k(x^c)$ to have its ij th entry given by

$$a_{ij}^k(x^c) = L_{g_j} L_f^{\gamma_i^k-1} h_i(x^c).$$

Let r_k be the normal rank of $A_k(x^c)$ in an open neighborhood of x_0^c . If $r_k = m$, stop.

Step ii: If $r_k < m$, define a square and nonsingular matrix $\hat{G}_k(x^c)$ such that the $(m - r_k)$ last columns of $\hat{A}_k(x^c) := A_k(x^c)\hat{G}_k(x^c)$ are identically zero. Moreover, this process can be carried out such that there exists r_k rows which the nonzero elements form an $r_k \times r_k$ nonsingular diagonal matrix. It is shown in [6] that $\hat{G}_k(x^c)$ always exists and is an analytic function of x^c .

There are r_k columns of $\hat{A}_k(x^c)$ with nonzero elements, out of which q_k columns have two or more nonzero elements. Design a permutation matrix P_k such that the first q_k columns of $\hat{A}_k(x^c)P_k$ have two or more elements nonzero, followed by the $(r_k - q_k)$ columns having only one nonzero element and finally the last $(m - r_k)$ columns having all zero elements. Denote $G_k(x^c) = \hat{G}_k(x^c)P_k$. Define an intermediate input \hat{u} by

$$\hat{u} = [G_k(x^c)]^{-1}u. \quad (7)$$

Step iii: The system \sum_k now is

$$\begin{aligned}\dot{x}^e &= f(x^e) + g(x^e)G_k(x^e)\hat{u} \\ y &= h(x^e).\end{aligned}$$

Add integrators in series with the first q_k inputs. This creates the new input vector \hat{u} of the form

$$\hat{u} = \begin{bmatrix} \dot{\hat{u}}_1 \\ \vdots \\ \dot{\hat{u}}_{q_k} \\ \hat{u}_{q_k+1} \\ \vdots \\ \hat{u}_m \end{bmatrix}. \quad (8)$$

Thus the new system after adding these integrators is

$$\begin{bmatrix} \dot{x}^e \\ \vdots \\ \dot{\hat{u}}_1 \\ \vdots \\ \dot{\hat{u}}_{q_k} \end{bmatrix} = \begin{bmatrix} f(x^e) + \sum_{i=1}^{q_k} \tilde{g}_i(x^e)\hat{u}_i \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} \sum_{i=q_k+1}^m \tilde{g}_i(x^e)\hat{u}_i \\ \vdots \\ \hat{u}_1 \\ \vdots \\ \hat{u}_{q_k} \end{bmatrix}$$

$$y = h(x^e) \quad (9)$$

where $\tilde{g}(x^e) = g(x^e)G_k(x^e)$. Call this system \sum_{k+1} .

Step iv: Go to Step i and resume the procedure with $k \leftarrow k+1$, new state variables $x^e \leftarrow \{x^e\} \cup \{\hat{u}_i\}_{i=1, \dots, q_k}$ and new input $u \leftarrow \hat{u}$. Let f, g, h still denote the extended f, g, h for notational simplicity. Let U still denote the open set of interest containing the equilibrium point of the extended system. \square

It has been shown by Descusse and Moog that if system (1) is right invertible and satisfies the accessibility rank condition (cf. [5, p. 86]) at x_0 , then the above algorithm converges in a finite number of steps, L , to an extended system \sum_L which is decouplable by static state feedback.

Note: At the end of i th iteration, the above algorithm adds q_i integrators ($q_i \leq r_i$) to the system \sum_i . Thus the dimension of x^e increases by q_i .

The Dynamic Extension algorithm (cf. [4]) and Singh's algorithm ([7], [15]) are similar, except that Singh's algorithm involves nonlinear transformations of the output space instead of the input space as in the Descusse-Moog algorithm.

The computation of the rank of $A_k(x^e)$ will be greatly simplified if $A_k(x^e)$ satisfies a regularity condition (see [16]), which guarantees that the normal rank of $A_k(x^e)$ is the same as the local rank for every $x^e \in U$. In this paper, we assume that \sum_0 is a regular system. If the given system is not regular, then one can use the procedure of [11] to get an approximate system that is regular and call this system \sum_0 .

Normal Form

Let us assume that the Descusse-Moog dynamic decoupling algorithm converges after L steps to a system of the form (6). Let us denote this extended system by \sum_e . Let (f^e, g^e, h^e) be the triple characterizing \sum_e , $x^e = (x, z) \in \mathbb{R}^{n+n_e}$ its state, u^e its input, and y^e its output. Let $x_0^e = (x_0, z_0)$ be the equilibrium point of interest. This system

\sum_e has a well defined vector relative degree $[\gamma_1^e, \dots, \gamma_m^e]$ at x_0^e . Let $\gamma^e := \sum_{i=1}^m \gamma_i^e$. We can construct a local change of coordinates $\phi(x^e) = (\xi, \eta)$ with $\xi = \text{col}(\xi^i)$, such that $\phi(x_0^e) = 0$, by choosing

$$\begin{aligned}\xi^i &= \text{col}(h_i^e(x^e), L_{f^e} h_i^e(x^e), \dots, L_{f^e}^{\gamma_i^e-1} h_i^e(x^e)) \\ &:= \text{col}(\xi_1^i, \xi_2^i, \dots, \xi_{\gamma_i^e}^i)\end{aligned} \quad (10)$$

and remaining $(n + n_e - \gamma^e)$ complementary coordinates η . In these coordinates, \sum_e takes the normal form (see [4])

$$\begin{aligned}\dot{\xi}_1^i &= \xi_2^i \\ &\vdots \\ \dot{\xi}_{\gamma_i^e-1}^i &= \xi_{\gamma_i^e}^i \\ \dot{\xi}_{\gamma_i^e}^i &= b_i^e(\xi, \eta) + \sum_{j=1}^m a_{ij}^e(\xi, \eta)u_j^e \\ \dot{\eta} &= q(\xi, \eta) + P(\xi, \eta)u^e \\ y_i^e &= \xi_1^i\end{aligned} \quad (11)$$

for $i = 1, \dots, m$, where

$$b_i^e(\xi, \eta) = L_{f^e}^{\gamma_i^e} h_i^e(\phi^{-1}(\xi, \eta)) \quad 1 \leq i \leq m$$

$$a_{ij}^e(\xi, \eta) = L_{g_j^e} L_{f^e}^{\gamma_i^e-1} h_i^e(\phi^{-1}(\xi, \eta)) \quad 1 \leq i, j \leq m.$$

The static state feedback that decouples the system \sum_e is given by

$$u^e = (A^e(\xi, \eta))^{-1}[-b^e(\xi, \eta) + v]. \quad (12)$$

The decoupling state feedback renders the η dynamics unobservable. The zero dynamics of system \sum_e are the dynamics of the η coordinates in the subspace $\xi = 0$ with the decoupling feedback law of (12) (with $v = 0$), i.e.,

$$\dot{\eta} = q(0, \eta) - P(0, \eta)[(A^e(0, \eta))^{-1}b^e(0, \eta)]. \quad (13)$$

For a detailed discussion of the zero dynamics and the transmission zeros of nonlinear systems we refer the reader to [17], [4, Chapter 6], [5, Chapter 11].

III. APPROXIMATE DYNAMIC DECOUPLING ALGORITHM

The difficulty in implementing the decoupling algorithm comes from the ill-conditioning of the decoupling matrix or from a situation in which the decoupling matrix may be nonsingular but close to singularity. To be precise, this occurs if the smallest singular value of $A^e(x^e)$ is smaller than a certain prespecified $\epsilon > 0$ for any $x \in U$. In this case, the algorithm calls for the inverse of an ill-conditioned matrix. In addition to the fact that the inverse is not numerically robust, it may cause large feedback gains in the controller and also cause the cancellation of far-off zeros (see [13, Section 4]). To alleviate these difficulties, we propose the following numerically robustified decoupling algorithm: while the algorithm appears to have numerical considerations in mind, it is, in fact, valuable for the reason that it does not cancel far off right-half plane zeros and may help control the magnitudes of the control inputs. To state the algorithm, recall a few basic facts and definitions about the numerical rank of a matrix.

Definition 1: A matrix $A \in \mathbb{R}^{n \times n}$ is said to have ϵ numerical rank r if

$$\inf \text{rank} \{B: \|B - A\| < \epsilon\} = r.$$

The norm in the above definition is the induced norm of the matrix induced by the Euclidean norm.

Thus, the numerical rank of a matrix is the lowest it can drop to in an ϵ neighborhood of the given matrix. In particular, if the matrix has $(n - r)$ of its singular values less than ϵ , then its numerical rank is r .

Approximate Dynamic Decoupling Algorithm. This algorithm starts at $k = 0$ with the given system Σ_0 having x_0 as its equilibrium point. We are given a threshold $\epsilon > 0$. Let the extended system at the end of iteration $(k - 1)$ be denoted by Σ_k having x^e as its state and x_k^e , the corresponding equilibrium point.

Step i: Compute the relative degrees of the outputs, namely, γ_i^k , $i = 1, \dots, m$, and the decoupling matrix $A_k(x^e)$. Let r_k be the normal rank of $A_k(x^e)$ in U . If $r_k = m$ and if the smallest singular value of $A_k(x^e)$ is greater than the threshold ϵ uniformly on U , stop.

Step ii: If all the nonzero singular values of $A_k(x^e)$ are less than ϵ uniformly on U , approximate $A_k(x^e)$ by a zero matrix. Go to Step i and recalculate the relative degrees γ_i^k with this approximation.

If $r_k = m$, go to Step iii, with $\hat{A}_k(x^e) = A_k(x^e)$ and $\hat{G}_k(x^e) = I_{m \times m}$.

Design a square, analytic and nonsingular matrix $\hat{G}_k(x^e)$ such that the last $m - r_k$ columns of $\hat{A}_k(x^e) := A_k(x^e)\hat{G}_k(x^e)$ are identically zero.

Step iii: If the smallest nonzero singular value of $A_k(x^e)$ is greater than the threshold ϵ , go to Step iv, with $\hat{A}_k(x^e) = \hat{A}_k(x^e)$, $G_k(x^e) = I_{m \times m}$ and $w_k = r_k$.

If the smallest nonzero singular value of $\hat{A}_k(x^e)$ is smaller than ϵ , then there exists a positive integer $w_k (< r_k)$ such that the ϵ rank of $\hat{A}_k(x^e)$ is w_k uniformly on U , i.e., $(r_k - w_k)$ nonzero singular values of $\hat{A}_k(x^e)$ are less than ϵ uniformly on U .

Design a square analytic nonsingular matrix $\hat{G}_k(x^e)^1$ such that

$$\begin{aligned} \hat{A}_k(x^e) &= \hat{A}_k(x^e)\hat{G}_k(x^e) \\ &= [a_1^k(x^e), \dots, a_{w_k}^k(x^e), \epsilon a_{w_k+1}^k(x^e), \dots, \\ &\quad \epsilon a_{r_k}^k(x^e), 0, \dots, 0]. \end{aligned} \quad (14)$$

Approximate the $r_k - w_k$ columns, which are small in norm, by identically zero columns. Go to Step iv with

$$\tilde{A}_k(x^e) = [a_1^k(x^e), \dots, a_{w_k}^k(x^e), 0, \dots, 0].$$

Note. The ϵ numerical rank of $\hat{A}_k(x^e)$ may not be constant over U . If the ϵ numerical rank passes through singularity at x_0^e , then the approximate algorithm will not work.² Throughout the analysis, we assume that the ϵ numerical rank of $\hat{A}_k(x^e)$ is constant in U .

¹ The proof of existence of such a matrix is given in the Appendix.

² The problem may be solved by reducing the value of ϵ . Since Σ_0 is a regular system (cf. [16]), the singularity does not exist for $\epsilon = 0$. However, reducing the value of ϵ defeats the purpose of the algorithm, by causing high gain solutions to the original decoupling problem.

Step iv: Out of w_k nonzero columns of $\tilde{A}_k(x^e)$, q_k columns will have two or more nonzero elements. Design a permutation matrix P_k such that the first q_k columns of $\tilde{A}_k(x^e)P_k(x^e)$ have two or more nonzero elements, followed by the $(w_k - q_k)$ columns having only one nonzero element and finally the last $(m - w_k)$ identically zero columns. Denote $G_k(x^e) = \tilde{G}_k(x^e)\tilde{G}_k(x^e)P_k$. Define an intermediate input by

$$\hat{u} = [G_k(x^e)]^{-1}u.$$

Step v: The system Σ_k now is

$$\begin{aligned} \dot{x}^e &= f(x^e) + g(x^e)G_k(x^e)\hat{u} \\ y &= h(x^e). \end{aligned}$$

Add an integrator in series with the first q_k inputs. This creates the new input vector \tilde{u} of the form

$$\tilde{u} = \begin{bmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_{q_k} \\ \hat{u}_{q_k+1} \\ \vdots \\ \hat{u}_m \end{bmatrix} = \begin{bmatrix} \tilde{u}_1 \\ \vdots \\ \tilde{u}_{q_k} \\ \hat{u}_{q_k+1} \\ \vdots \\ \hat{u}_m \end{bmatrix}. \quad (15)$$

Thus the new system after adding these integrators is

$$\begin{aligned} \begin{bmatrix} x^e \\ \vdots \\ \hat{u}_1 \\ \vdots \\ \hat{u}_{q_k} \end{bmatrix} &= \begin{bmatrix} f(x^e) = \sum_{i=1}^{q_k} \hat{q}_i(x^e)\hat{u}_i \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{m-q_k} \hat{q}_i(x^e)\hat{u}_i \\ \vdots \\ \hat{u}_1 \\ \vdots \\ u_{q_k} \end{bmatrix} \\ y &= h(x^e) \end{aligned} \quad (16)$$

where $q(x^e) = g(x^e)G_k(x^e)$. Call this system Σ_{k+1} .

Step vi: Return to Step i and resume the procedure with $k \leftarrow k + 1$, new state variables $x^e \leftarrow \{x^e\} \cup \{\hat{u}_i\}_{i=1, \dots, q_k}$ and new input $u \leftarrow \tilde{u}$. Let f, g, h still denote the extended f, g, h for notational simplicity. Let U still denote the open set containing the equilibrium point of the extended system. \square

Let us assume that the approximate dynamic decoupling algorithm converges after L steps to a system of the form of (6). Let us denote this extended system by $\tilde{\Sigma}_e$. Let $(\tilde{f}^e, \tilde{g}^e, \tilde{h}^e)$ be the triple characterizing $\tilde{\Sigma}_e$, $\tilde{x}^e = (x, z)$ its state, \tilde{u}^e its input, and \tilde{y}^e its output. Let $\tilde{x}_0^e = (x_0, z_0)$ be the equilibrium point of interest. The system $\tilde{\Sigma}_e$ has a well-defined vector relative degree $[\tilde{\gamma}_1^e, \dots, \tilde{\gamma}_l^e]$ at \tilde{x}_0^e . Thus we can construct a local diffeomorphism of the form of (10) such that $\tilde{\Sigma}_e$ will be transformed into the normal form $(\tilde{\xi}, \tilde{\eta})$ coordinates given by

$$\begin{aligned} \tilde{\xi}^e &= \text{col}(\tilde{h}_1^e(x^e), L_{\tilde{f}_e} \tilde{h}_1^e(\tilde{x}^e), \dots, L_{\tilde{f}_e}^{\tilde{\gamma}_1^e-1} \tilde{h}_1^e(\tilde{x}^e)) \\ &= \text{col}(\tilde{\xi}_1^e, \dots, \tilde{\xi}_{\tilde{\gamma}_1^e}^e) \end{aligned} \quad (17)$$

and remaining $(n + \tilde{n}_e - \sum_{i=1}^m \tilde{\gamma}_i^e)$ complementary coordinates $\tilde{\eta}$. In the course of the approximate dynamic decoupling algorithm, we have neglected order ϵ terms at each iteration of

algorithm. Thus the original or exact system in the normal coordinates of \sum_e will be

$$\begin{aligned}\dot{\tilde{\xi}}_1^i &= \tilde{\xi}_2^i \\ &\vdots \\ \dot{\tilde{\xi}}_{\gamma_i^0-1}^i &= \tilde{\xi}_{\gamma_i^0}^i \\ \dot{\tilde{\xi}}_{\gamma_i^0}^i &= \tilde{\xi}_{\gamma_i^0+1}^i + \epsilon \sum_{j=1}^m (\beta_{\gamma_i^0}^i)_j \tilde{u}_j^e \\ &\vdots \\ \dot{\tilde{\xi}}_{\gamma_i^e-1}^i &= \tilde{\xi}_{\gamma_i^e}^i + \epsilon \sum_{j=1}^m (\beta_{\gamma_i^e-1}^i)_j \tilde{u}_j^e \\ \dot{\tilde{\xi}}_{\gamma_i^e}^i &= \tilde{b}_i^e(\tilde{\xi}, \tilde{\eta}) + \sum_{j=1}^m \tilde{a}_{ij}^e(\tilde{\xi}, \tilde{\eta}) \tilde{u}_j^e \\ \dot{\tilde{\eta}} &= q(\tilde{\xi}, \tilde{\eta}) + P(\tilde{\xi}, \tilde{\eta}) \tilde{u}^e \\ \tilde{y}_i^e &= \tilde{\xi}_1^i\end{aligned}\quad (18)$$

for $i = 1, \dots, m$, where

$$\tilde{b}_i^e(\tilde{\xi}, \tilde{\eta}) = L_{f^e}^{\gamma_i^e} \tilde{h}_i^e(\phi^{-1}(\tilde{\xi}, \tilde{\eta})) \quad 1 \leq i \leq m$$

$$\tilde{a}_{ij}^e(\tilde{\xi}, \tilde{\eta}) = L_{g_j^e}^{\gamma_i^e-1} \tilde{h}_i^e(\phi^{-1}(\tilde{\xi}, \tilde{\eta})) \quad 1 \leq i, j \leq m.$$

Note: If we substitute $\epsilon = 0$ in (18), then we get the representation of the system \sum_e in its normal form, coordinates $(\tilde{\xi}, \tilde{\eta})$. The static state feedback that decouples the system \sum_e is given by

$$\tilde{u}^e = (\tilde{A}^e(\tilde{\xi}, \tilde{\eta}))^{-1} [-\tilde{b}^e(\tilde{\xi}, \tilde{\eta}) + v]. \quad (19)$$

If we compare the approximate and exact decoupling algorithm, we see that

- If the exact decoupling algorithm converges with a decoupling matrix $A^e(x^e)$ whose smallest singular value is greater than ϵ uniformly on U , then the approximate decoupling algorithm yields the same result.
- If, on the other hand, $A^e(x^e)$ obtained by exact decoupling algorithm has its smallest singular value of the order of ϵ uniformly on U , then the decoupling control law (12) will have terms of the order of $1/\epsilon$ that will result in a high gain controller that may not be practically feasible if the actuators have saturation limits.

The approximate decoupling algorithm is of use if we can answer the following two questions:

- 1) If the given system (1) is right invertible and locally accessible (i.e., the exact decoupling algorithm converges in L steps), then does the approximate decoupling algorithm converge in a finite number of steps?
- 2) If the answer to the first question is yes, then how much error do we introduce in tracking the reference outputs by using the approximate decoupling and tracking law instead of the exact one?

These questions will be answered in the following sections.

IV. CONVERGENCE OF APPROXIMATE DECOUPLING ALGORITHM

The approximate decoupling algorithm is based on the Descusse and Moog dynamic decoupling algorithm. Consequently, the convergence properties of the approximate decoupling algorithm will be compared with that of the Descusse and Moog algorithm. Ideally the approximate decoupling algorithm should preserve the convergence properties of the Descusse and Moog algorithm.

The algorithm of Descusse and Moog adds dynamics to the given system \sum_0 , to get an extended system \sum_L which is decouplable by static state feedback. The following theorem [6], relates this notion to the nonsingularity of the decoupling matrix $A_L(x)$.

Theorem 1: System \sum of the form of (1) is decouplable by static state feedback, if, the decoupling matrix for \sum is invertible.

The approximate decoupling algorithm converges to an extended system that is robustly decouplable by static state feedback. This notion is defined as follows.

Definition 2: An $m \times m$ matrix $A(x)$ is ϵ robustly invertible in an open set U with respect to a threshold ϵ if the ϵ numerical rank of $A(x)$ is m uniformly on U .

Definition 3: A system \sum of the form of (1) is ϵ robustly decouplable by static state feedback with respect to a threshold ϵ , if, \sum is decouplable by static state feedback and the decoupling matrix $A(x)$ is robustly invertible with respect to ϵ .

The following lemma considers the effect of one iteration of the approximate decoupling algorithm on a system that is decouplable by using the Descusse and Moog algorithm.

Lemma 1: Suppose that the Descusse and Moog algorithm converges for a system \sum_k of the form of (1). Apply the approximate decoupling algorithm. After one iteration, we get the extended system \sum_{k+1} . One of the following is true for \sum_{k+1} :

- 1) $\gamma_i^{k+1} = \infty$ for some i .
- 2) $A_{k+1}(x)$ is singular and the Descusse and Moog decoupling algorithm does not converge for \sum_{k+1} .
- 3) $\gamma^{k+1} = n_{\sum_{k+1}}$ and $A_{k+1}(x)$ is not robustly invertible.
- 4) $A_{k+1}(x)$ is singular, but the Descusse and Moog algorithm converges for \sum_{k+1} .
- 5) $A_{k+1}(x)$ is nonsingular, not robustly invertible and $\gamma^{k+1} < n_{\sum_{k+1}}$.
- 6) $A_{k+1}(x)$ is robustly invertible.

where $\gamma^{k+1} = \sum_{i=1}^m \gamma_i^{k+1}$ and $n_{\sum_{k+1}}$ = the dimension of state space of \sum_{k+1} .

Proof: This is a list of all the cases after application of one step of the approximate decoupling algorithm. It is easy to check that the list exhausts all the possibilities. \square

We will analyze each of the above cases in detail to understand why in some cases the approximate decoupling algorithm may not converge.

While applying the Descusse and Moog algorithm, we differentiate each output until at least one input appears on the right hand side. Some of the inputs show up earlier than others making the decoupling matrix singular. Integrators are

added in front of these inputs to delay their appearance for at least one more step of differentiation.

In this process, at a particular step, some of the inputs might be weakly connected to the outputs, i.e., the functions multiplying them are smaller than the threshold ϵ uniformly in U . These functions are approximated by zero in the approximate decoupling algorithm. This modification in the original Descusse and Moog algorithm might make the resulting systems noninvertible. The classification of various cases in the previous lemma helps detect such systems in the following manner:

Analysis of Lemma 1. Recall that γ_i^k are the relative degrees of the outputs of Σ_k . We have

$$\begin{bmatrix} y_1^{\gamma_1^k} \\ \vdots \\ y_m^{\gamma_m^k} \end{bmatrix} = b_k(x) + A_k(x)u$$

rank of $A_k(x) = r_k$. At the end of Step ii, $(m - r_k)$ columns of $\hat{A}_k(x)$ are identically zero and there is at least one nonzero element in each row of $\hat{A}_k(x)$.

At the end of Step iv, some of the rows of $A_k(x)G_k(x)$ might be identically zero because of the approximation of $(r_k - w_k)$ columns by zero. The number of such rows is less than or equal to $(r_k - w_k)$. Let us denote the set of outputs corresponding to these rows by \bar{Y}_k .

$(w_k - q_k)$ columns of $A_k(x)G_k(x)$ have only one nonzero element. By construction, these nonzero elements will be in $(w_k - q_k)$ different rows. Let us denote the $(w_k - q_k)$ outputs corresponding to these rows by \hat{Y}_k . The remaining outputs will be denoted by \tilde{Y}_k .

The outputs of Σ_k do not change in the process. Thus for Σ_{k+1} , we have

- $\gamma_i^{k+1} = \gamma_i^k, \forall i \text{ s.t. } y_i \in \hat{Y}_k$
- $\gamma_i^{k+1} = \gamma_i^k + 1, \forall i \text{ s.t. } y_i \in \tilde{Y}_k$
- $\gamma_i^{k+1} \geq \gamma_i^k + 1, \forall i \text{ s.t. } y_i \in \bar{Y}_k$

Case 1: If the only nonzero entries in the i th row are in the j th column, then when the j th column of $A_k(x)G_k(x)$ is approximated by zero, the i th row is made identically zero. y_i is only affected by u_j , and after the approximation, u_j never appears on the right hand side again. This makes $\gamma_i^{k+1} = \infty$.

Case 2: The situation here is that Case 1 recurs, but not immediately at the end of the $k - 1$ st step but later in the algorithm. Thus the system Σ_k loses invertibility because of the approximation. Some rows of $A_{k+l}(x)$ will be dependent for all $l > 0$. This case can also be avoided by normalizing the j th input.

In the course of the approximate decoupling algorithm, Case 2 goes unnoticed until you reach Case 3. Thus the reason for Case 3 is in fact the occurrence of Case 2 during one of the previous iterations.

If the approximate decoupling algorithm converges to a system Σ_L that can be decoupled by static state feedback, the outputs y_i and their respective derivatives up to the order of $(\gamma_i^L - 1)$ qualify as a partial change of coordinates. In the normal form notation these are the ξ coordinates.

During each iteration of the approximate decoupling algorithm, the state space dimension of Σ_k is extended by

q_k whereas at least $(m - w_k + q_k)$ new ξ coordinates are introduced. The difference between the state space dimension of Σ_{k+1} and the dimension of ξ coordinates decreases by $(m - w_k)$ during each iteration of the approximate decoupling algorithm.

Case 3: If we go through $(k + 1)$ th iteration, dimension of ξ coordinates will exceed the dimension of the state space of Σ_{k+2} . Thus we cannot proceed further.

Cases 4, 5, and 6 lead towards the convergence of approximate decoupling algorithm. \square

Definition 4: Suppose Σ_0 satisfies the hypothesis of Descusse and Moog algorithm. Apply the approximate decoupling algorithm, for given $\epsilon > 0$. Let $k \geq 0$ be the smallest integer such that either case 1, 2, or 3 of the previous lemma is true for Σ_{k+1} . Then the system Σ_0 is said to be ϵ -unnormalized.

Theorem 2. Suppose Σ_0 satisfies the hypothesis of the Descusse and Moog dynamic decoupling algorithm. Apply the approximate decoupling algorithm for a given $\epsilon > 0$. Then one of the following is true.

- The approximate decoupling algorithm converges in a finite number of steps.
- The system Σ_0 is ϵ -unnormalized.

Proof: During each step of the approximate decoupling algorithm, the difference between the state space dimension and the dimension of ξ coordinates decreases by $(m - w_k)$. Thus, if the first three cases of the previous lemma are avoided during each iteration, the algorithm has to converge in a finite number of steps.

The discussion following the previous lemma shows that the first three cases correspond to the underlying system being unnormalized. \square

In general, it is not possible, *a priori*, to find out whether a given system is normalized or not. If the approximate decoupling algorithm does not converge in n steps, then the system is unnormalized, provided it was decouplable by using the exact Descusse and Moog dynamic decoupling algorithm.

The design parameter ϵ is an input to the approximate decoupling algorithm. If the approximate algorithm does not converge for a given ϵ , because of the system being ϵ unnormalized, then the value of ϵ can be reduced. This might result in convergence of the algorithm but the original problem of numerical robustness and high gain solutions still exists. In the limit that $\epsilon \rightarrow 0$, one recovers the Descusse–Moog algorithm. For $\epsilon = 0$, the ϵ unnormalized systems are noninvertible. Thus given an invertible system that satisfies the accessibility rank condition, one can always find an ϵ so that the approximate algorithm converges. In particular, the algorithm converges for the specific examples of [1], [2], [3].

A. Multiple Time Scale Zero Dynamics

Since the zero dynamics of a system does not change by addition of integrators to its input channels [4, p. 389], by input space transformations or by state feedback, the zero dynamics of Σ_0 is same as that of Σ_L , where Σ_L is the extended system at the end of Descusse and Moog algorithm. The next lemma compares the zero dynamics of Σ_L and Σ_L where $\tilde{\Sigma}_L$ is the extended system at the end of approximate

coupling algorithm. Note that the approximate decoupling algorithm cannot converge in fewer steps than the Descusse and Moog algorithm.

Lemma 2: Suppose Σ_0 is right invertible (cf. [6]) and satisfies the strong accessibility rank condition (cf. [5, p. 86]) at r_0 . Suppose that the approximate decoupling algorithm converges for this system in exactly the same number of steps as the Descusse and Moog algorithm. Let Σ_L and $\tilde{\Sigma}_L$ be the system at the end of the Descusse and Moog algorithm and the approximate decoupling algorithms respectively. Then

$$\dim(\eta_{\tilde{\Sigma}_L}) \leq \dim(\eta_{\Sigma_L})$$

where η denotes the zero dynamics coordinates.

Proof: During each iteration, the difference between state space coordinates and the ξ coordinates decreases by $(m - w_k)$ for approximate decoupling algorithm and by $(m - r_k)$ for the Descusse and Moog decoupling algorithm. After L steps when both the algorithms converge, the zero dynamics dimension is given by the difference between the state space dimension of the extended system and the dimension of the ξ coordinates. Thus

$$\begin{aligned} \dim(\eta_{\tilde{\Sigma}_L}) &= \dim(\eta_{\Sigma_L}) - \sum_{i=0}^{L-1} (r_i - w_i) \\ &\leq \dim(\eta_{\Sigma_L}). \end{aligned} \quad \square$$

Thus, in general, the zero dynamics of the extended system at the end of the approximate decoupling algorithm (i.e., $\tilde{\Sigma}_L$) will have smaller dimension than that of Σ_L . We would like to investigate the relationship between $\eta_{\tilde{\Sigma}_L}$ and η_{Σ_L} .

Recent results in the area of singularly perturbed zero dynamics of nonlinear systems ([12], [13]) lead us to the following conclusion: under some suitable technical hypotheses, the zero dynamics of Σ_L can be decomposed into two or more time scales using singular perturbation theory (cf. [18]). The slow or reduced system is described by the zero dynamics of $\tilde{\Sigma}_L$ and the dynamics that were neglected during the process of approximation constitutes the faster time scale or boundary layer subsystem. In [13], the authors have proved the above conclusion for a restricted class of two-input/two-output systems. The full details and the technical hypotheses needed to guarantee the existence of singularly perturbed zero dynamics for this general class of systems remain to be worked out. We will conclude this section with the following remarks.

The approximate decoupling algorithm creates an extended system that does not include the far off zeros of the original system. Since the static state feedback that achieves decoupling of Σ_L is a pole zero cancellation law, we do not cancel the far off zeros of Σ_0 in the case of the approximate decoupling algorithm. The cancellation of these far off zeros requires a large control effort resulting in a high gain controller. If these far off zeros are unstable, then their cancellation makes the closed-loop system unstable. These systems are referred to as slightly nonminimum phase systems in [2]. Application of the approximate decoupling algorithm to slightly nonminimum phase systems results in a stable closed-loop system.

V. APPROXIMATE ASYMPTOTIC TRACKING

Input-output decoupling is closely related to tracking of reference trajectories by the outputs of a MIMO nonlinear system. If the desired trajectories to be tracked fall into a restricted class of functions, say constants or sinusoids with a finite spectrum, then we can use nonlinear regulator theory (see [4, Chapter 7], [19], [20]). If the class of desired trajectories is more general, for example, functions that are N times continuously differentiable but otherwise arbitrary, then according to [21] the decoupling controller forms an inner loop of the overall tracking controller. If the given system is not robustly decouplable by using exact decoupling algorithms, then we have to use the approximate decoupling feedback. This section considers the effects of approximate decoupling on the performance and stability of the overall tracking controller.

Let us assume that the approximate decoupling algorithm converges for Σ_0 giving us the approximate extended system $\tilde{\Sigma}_L$. Equations (18) with $\epsilon = 0$ represent $\tilde{\Sigma}_L$ in its normal form $(\tilde{\xi}, \tilde{\eta})$ coordinates. If the objective of the controller is to track the desired reference trajectory $y_d(t) = [y_{d_1}(t), \dots, y_{d_m}(t)]^T$ which is smooth and bounded with bounded derivatives, we design the control input \hat{u}_L to be

$$\hat{u}^c = (\hat{A}'(\tilde{\xi}, \tilde{\eta}))^{-1}[-b'(\tilde{\xi}, \tilde{\eta}) + v]$$

$$\begin{aligned} v &= \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} \\ &= \begin{bmatrix} y_{d_1}^{\gamma_1^i} + \sigma_{\gamma_1^i-1}^1 (y_{d_1}^{\gamma_1^i-1} - \xi_{\gamma_1^i-1}^1) + \dots + \sigma_0^1 (y_{d_1} - \xi_1^1) \\ \vdots \\ y_{d_m}^{\gamma_m^i} + \sigma_{\gamma_m^i-1}^m (y_{d_m}^{\gamma_m^i-1} - \xi_{\gamma_m^i-1}^m) + \dots + \sigma_0^m (y_{d_m} - \xi_1^m) \end{bmatrix} \end{aligned} \quad (20)$$

where $(s^{\gamma_i^i} + \sigma_{\gamma_i^i-1}^i s^{\gamma_i^i-1} + \dots + \sigma_0^i)$ is a Hurwitz polynomial for $i = 1, \dots, m$.

Let us define the tracking errors to be

$$e_i^t := \tilde{\xi}_i^t - y_{d_i}, \quad 1 \leq i \leq m. \quad (21)$$

Let us define the error coordinates for system $\tilde{\Sigma}_L$ to be

$$\begin{bmatrix} e_1^t \\ \vdots \\ e_m^t \end{bmatrix} = \begin{bmatrix} \tilde{\xi}_1^t \\ \vdots \\ \xi_{\gamma_i^i-1}^t \end{bmatrix} - \begin{bmatrix} y_{d_1} \\ \vdots \\ y_{d_i}^{\gamma_i^i-1} \end{bmatrix} \quad 1 \leq i \leq m. \quad (22)$$

Thus the system $\tilde{\Sigma}_L$ with the feedback (20) can be expressed in $(e, \tilde{\eta})$ coordinates by

$$\dot{e}^t = \bar{A}' e^t \quad i = 1, \dots, m$$

$$\dot{\tilde{\eta}} = q(\tilde{\xi}, \tilde{\eta}) + P(\tilde{\xi}, \tilde{\eta}) \hat{u}^c(\tilde{\xi}, \tilde{\eta}, v) \quad (23)$$

where $\bar{A}' \in \mathbb{R}^{\tilde{\gamma}_i^i \times \tilde{\gamma}_i^i}$ given by

$$\bar{A}' = \begin{bmatrix} 0 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & 0 & 1 \\ -\sigma_0^i & -\sigma_1^i & \dots & -\sigma_{\gamma_i^i-1}^i \end{bmatrix}.$$

It can be shown (e.g., see [2]) that if

- The reference trajectory and its derivatives are bounded and small enough.
- Zero dynamics of (23) (i.e., the equilibrium point $\tilde{\eta}_0 = 0$ of the system)

$$\dot{\tilde{\eta}} = q(0, \tilde{\eta}) + P(0, \tilde{\eta})\tilde{u}^e(0, \tilde{\eta}, 0) \quad (24)$$

is exponentially stable.

- $q(\tilde{\xi}, \tilde{\eta}) + P(\tilde{\xi}, \tilde{\eta})\tilde{u}^e(\tilde{\xi}, \tilde{\eta}, v)$ is locally Lipschitz continuous in $\tilde{\xi}, \tilde{\eta}$ then $\lim_{t \rightarrow \infty} e_1^i(t) = 0 \forall i$, and the states $\tilde{\xi}, \tilde{\eta}$ remain bounded.

The controller of (20) is designed for the approximate extended system \sum_e . If we apply this controller to the exact system, we get the system equations in the $(e, \tilde{\eta})$ coordinates given by

$$\begin{aligned} \dot{e}^i &= \bar{A}^i e^i + \epsilon \beta^i(x^e) \tilde{u}^e(x^e) \quad 1 \leq i \leq m \\ \dot{\tilde{\eta}} &= q(\tilde{\xi}, \tilde{\eta}) + P(\tilde{\xi}, \tilde{\eta})\tilde{u}^e(\tilde{\xi}, \tilde{\eta}, v) \end{aligned} \quad (25)$$

where

$$\beta^i = \begin{bmatrix} 0_{1 \times m} \\ \vdots \\ 0_{1 \times m} \\ \beta_{\tilde{\eta}_1^0}^i \\ \vdots \\ \beta_{\tilde{\eta}_i^0}^i \\ \vdots \\ \beta_{\tilde{\eta}_i^0}^i \\ 0_{1 \times m} \end{bmatrix}$$

represents the dynamics that was neglected during the approximate decoupling algorithm. Each β_j^i is $1 \times m$ row vector of functions of x , the first w_0 elements of which are identically zero.

The tracking control law (20) was designed for a system of the form (25) with $\epsilon = 0$. The following theorem shows that it works for the approximate system with nonzero ϵ as well. This theorem is motivated by and is similar to the one for slightly nonminimum phase systems as in [2].

Theorem 3: If

- Zero dynamics of system (25) (i.e., the equilibrium point $\tilde{\eta}_0 = 0$ of (24)) is exponentially stable in a neighborhood U .
- The functions $\beta^i(x^e) \tilde{u}^e(x^e)$ are locally Lipschitz continuous in a neighborhood U with $\beta^i(x_0^e) \tilde{u}^e(x_0^e) = 0 \forall i = 1, \dots, m$.
- $q(\tilde{\xi}, \tilde{\eta}) + P(\tilde{\xi}, \tilde{\eta})\tilde{u}^e(\tilde{\xi}, \tilde{\eta}, v)$ is locally Lipschitz continuous in $\tilde{\xi}, \tilde{\eta}, v$.

Then for ϵ sufficiently small and for desired trajectories with derivatives small enough, the states of system (25) are bounded and the tracking errors satisfy

$$\|e_1^i\| = \|\tilde{\xi}_1^i - y_d\| \leq K\epsilon$$

for some $K < \infty$.

Note: Since we know that f, g, h are smooth functions of x to start with, the functions $\beta^i(x^e) \tilde{u}^e(x^e)$ will be locally Lipschitz so long as the matrix $G_k(x^e)$ is a smooth function of x^e at each iteration of the approximation decoupling algorithm.

Proof: From (22) and the fact that the desired trajectory and its derivatives are bounded (by b_d), we get

$$\|\tilde{\xi}\| \leq \|e\| + b_d. \quad (26)$$

The transformation that transforms \sum_e into $(\tilde{\xi}, \tilde{\eta})$ coordinates is a diffeomorphism, thus there exists $l_x > 0$ such that

$$\|x\| \leq l_x(\|\tilde{\xi}\| + \|\tilde{\eta}\|). \quad (27)$$

As the functions $\beta^i(x^e) \tilde{u}^e(x^e)$ are locally Lipschitz continuous with $\beta^i(x_0^e) \tilde{u}^e(x_0^e) = 0$, there exists a positive constant l_β such that

$$\|2P\beta(x^e) \tilde{u}^e(x^e)\| \leq l_\beta \|x^e\| \quad (28)$$

where $\beta(x^e)$ is the block diagonal matrix with $\beta^i(x^e)$ being its diagonal blocks. Since the zero dynamics

$$\dot{\tilde{\eta}} = q(0, \tilde{\eta}) + P(0, \tilde{\eta})\tilde{u}^e(0, \tilde{\eta}, 0) \quad (29)$$

is exponentially stable, by a converse Lyapunov theorem [22] there exists, $\bar{v}(\tilde{\eta})$ and positive constants k_1, k_2, k_3, k_4 such that

$$k_1 \|\tilde{\eta}\|^2 \leq \bar{v}(\tilde{\eta}) \leq k_2 \|\tilde{\eta}\|^2$$

$$\frac{\partial \bar{v}}{\partial \tilde{\eta}} [q(0, \tilde{\eta}) + P(0, \tilde{\eta})\tilde{u}^e(0, \tilde{\eta}, 0)] \leq -k_3 \|\tilde{\eta}\|^2$$

$$\left\| \frac{\partial \bar{v}}{\partial \tilde{\eta}} \right\| \leq k_4 \|\tilde{\eta}\|.$$

Thus from the exponential stability of zero dynamics and the Lipschitz continuity of $q + Pu^k$, we get

$$\begin{aligned} \dot{v} &= \frac{\partial \bar{v}}{\partial \tilde{\eta}} [q(\tilde{\xi}, \tilde{\eta}) + P(\tilde{\xi}, \tilde{\eta})u^k(\tilde{\xi}, \tilde{\eta}, v)] \\ &= \frac{\partial \bar{v}}{\partial \tilde{\eta}} [q(0, \tilde{\eta}) + P(0, \tilde{\eta})u^k(0, \tilde{\eta}, 0)] \\ &\quad + \frac{\partial \bar{v}}{\partial \tilde{\eta}} [q(\tilde{\xi}, \tilde{\eta}) + P(\tilde{\xi}, \tilde{\eta})u^k(\tilde{\xi}, \tilde{\eta}, v) \\ &\quad - \{q(0, \tilde{\eta}) + P(0, \tilde{\eta})u^k(0, \tilde{\eta}, 0)\}] \\ &\leq -k_3 \|\tilde{\eta}\|^2 + k_4 \|\tilde{\eta}\| l_\eta (\|\tilde{\xi}\| + \|v\|). \end{aligned}$$

To show the states of (25) remain bounded, let

$$V(e, \tilde{\eta}) = e^T \bar{P} e + \psi \bar{v}(\tilde{\eta})$$

be a Lyapunov function for the system (25) where $\bar{P} > 0$ satisfies the following Lyapunov equation

$$\bar{A}^T \bar{P} + \bar{P} \bar{A} = -I$$

and ψ is a positive constant to be specified later. Then

$$\begin{aligned} \dot{V} &= -\|e\|^2 + 2\epsilon e^T \bar{P} \beta(x) u^k(x) + \psi \frac{\partial \bar{v}}{\partial \tilde{\eta}} \dot{\tilde{\eta}} \\ &\leq -\|e\|^2 + \epsilon \|e\| l_\beta \|x\| \\ &\quad + \psi [-k_3 \|\tilde{\eta}\|^2 + k_4 \|\tilde{\eta}\| l_\eta (\|\tilde{\xi}\| + \|v\|)] \\ &\leq -\|e\|^2 + \epsilon \|e\| l_\beta l_x (\|e\| + b_d + \|\tilde{\eta}\|) + \psi \\ &\quad \cdot [-k_3 \|\tilde{\eta}\|^2 + k_4 \|\tilde{\eta}\| l_\eta (\|e\| + b_d + l_v (\|e\| + b_d))] \\ &\leq -\left(\frac{\|e\|}{2} - \epsilon l_\beta l_x b_d\right)^2 + (\epsilon l_\beta l_x b_d)^2 \end{aligned}$$

$$\begin{aligned}
 & - \left(\frac{\|e\|}{2} - (\epsilon l_\beta l_x + \psi k_4 l_\eta (1 + l_v)) \|\tilde{\eta}\| \right)^2 \\
 & + (\epsilon l_\beta l_x + \psi k_4 l_\eta (1 + l_v))^2 \|\tilde{\eta}\|^2 \\
 & - \psi k_3 \left(\frac{\|\tilde{\eta}\|}{2} - \frac{k_4 l_\eta b_d (1 + l_v)}{k_3} \right)^2 \\
 & + \psi \frac{(k_4 l_\eta b_d (1 + l_v))^2}{k_3} \\
 & - \left(\frac{1}{2} - \epsilon l_\beta l_x \right) \|e\|^2 - \frac{3}{4} \psi k_3 \|\tilde{\eta}\|^2 \\
 & \leq - \left(\frac{1}{2} - \epsilon l_\beta l_x \right) \|e\|^2 \\
 & - \left(\frac{3}{4} \psi k_3 - (\epsilon l_\beta l_x + \psi k_4 l_\eta (1 + l_v))^2 \right) \|\tilde{\eta}\|^2 \\
 & + (\epsilon l_\beta l_x b_d)^2 + \psi \frac{(k_4 l_\eta b_d (1 + l_v))^2}{k_3}.
 \end{aligned}$$

Let

$$\tilde{\psi} = \frac{k_3}{4(l_\beta l_x + k_4 l_\eta (1 + l_v))^2}.$$

Then, for all $\psi \leq \tilde{\psi}$ and all $\epsilon \leq \min(\tilde{\psi}, 1/4l_\beta l_x)$, we have

$$\dot{V} \leq -\frac{\|e\|}{4} - \frac{\psi k_3 \|\tilde{\eta}\|^2}{2} + (\epsilon l_\beta l_x b_d)^2 + \psi \frac{(k_4 l_\eta b_d (1 + l_v))^2}{k_3}. \quad (30)$$

Thus $\dot{V} < 0$ whenever $\|\tilde{\eta}\|$ or $\|e\|$ is large. This implies that $\|\tilde{\eta}\|$ and $\|e\|$ and also, $\|\xi\|$ and $\|x\|$ are bounded. The above analysis shows that if we choose the initial condition sufficiently close to x_0^c and b_d sufficiently small, we can guarantee that the states will remain in U . Using the boundedness of x and the continuity of $\beta^i(x)\tilde{u}^c(x)$, we see that

$$\dot{e}^i(x) = \bar{A}^i e^i + \epsilon \beta^i(x) \tilde{u}^c(x)$$

are m SISO exponentially stable linear systems driven by order ϵ input. Thus we conclude that the tracking errors e^i converge to a ball of order ϵ . \square

VI. CONCLUSION

A numerically robust algorithm for input-output decoupling of nonlinear dynamical systems has been proposed. This algorithm provides low gain, practically implementable controllers that do not cancel far off zeros. The use of this algorithm for a slightly nonminimum phase system (i.e., one that has far off right-half plane zeros), results in an overall stable closed-loop system. It is shown that the tracking controllers constructed by using this approximate decoupling algorithm result in bounded tracking with stability. Controllers based on this theory already exist for a few specific examples in the literature and this paper can be thought of as an attempt to formalize the techniques used in those particular examples. Detailed calculations to establish one-to-one mapping between the neglected dynamics and the formal structure at infinity are still to be worked out. The ongoing work on a CAD package called AP_LIN ([23]) based on this approach will help automate the application of

this theory. Although this approximate algorithm is based on the Descusse-Moog dynamic decoupling algorithm, a similar algorithm based on the dynamic extension algorithm (see [1, Chapter 7]) can be worked out in similar fashion.

APPENDIX

EXISTENCE OF AN ANALYTIC MATRIX $G_k(x)$

The proof of existence of $\hat{G}_k(x)$ is given by Descusse and Moog in [6]. Thus we have a matrix $\hat{A}_k(x)$ whose ϵ numerical rank is w_k and the last $m - r_k$ columns are identically zero.

From the definition of ϵ numerical rank of $\hat{A}_k(x)$, it is possible to find a $w_k \times w_k$ minor, say, $\Delta_k(x)$, such that all the singular values of $\Delta_k(x)$ are bigger than ϵ in U . Without loss of generality, we assume that $\Delta_k(x)$ is the block formed by the first w_k rows and the first w_k columns of $\hat{A}_k(x)$, and let $\delta(x)$ represent its determinant.

By definition, any minor of $\hat{A}_k(x)$ having size bigger than w_k , will have at least one singular value smaller than ϵ uniformly in U . Thus the determinant of this minor will be of the order of $\epsilon \times \delta(x)$. Thus we get

$$\det \begin{bmatrix} & & & \vdots & A_{1j}(x) \\ & & & \vdots & \vdots \\ & \Delta_k(x) & & \vdots & \vdots \\ & & & \vdots & A_{w_k j}(x) \\ \dots & \dots & \dots & \dots & \dots \\ A_{i1}(x) & \dots & A_{iw_k}(x) & \vdots & A_{ij}(x) \end{bmatrix} = \epsilon \delta(x)$$

$$\forall i \in \{w_k + 1, \dots, m\}, \quad \forall j \in \{w_k + 1, \dots, r_k\}. \quad (31)$$

Consider the top left $r_k \times r_k$ block of $\hat{A}_k(x)$. We get

$$\delta(x) A_{ij}(x) + \sum_{i=1}^{r_k} \lambda_i(x) A_{ii}(x) = \text{order } \epsilon \times \delta(x),$$

$$\forall i \in \{w_k + 1, \dots, m\}, \quad \forall j \in \{1, \dots, r_k\} \quad (32)$$

where $\lambda_i(x)$ is the cofactor of $A_{ii}(x)$, calculated with respect to the top left $r_k \times r_k$ block of $\hat{A}_k(x)$.

Define the elementary column operation by

$$g_j(x) := \begin{bmatrix} 1 & & & \lambda_1(x) \\ & \ddots & & \vdots \\ & & 1 & \lambda_{r_k}(x) \\ & & & 0 \\ & & & \vdots & 0 \\ & & & 1 & 0 \\ 0 & & & & \delta(x) \\ & & & & 0 \\ & & & & 1 \\ & & & & \vdots \\ & & & & 0 & 1 \end{bmatrix}$$

where $\delta(x)$ is in the j th row and column. The first $r_k - 1$ elements in the j th column of $\hat{A}_k(x)g_j(x)$ will be zero. The r_k^{th} element will be the determinant given by (32), which is

of the order of ϵ . The rest of the elements in this column must be of the order of ϵ , or else there will be a minor of $\hat{A}_k(x)g_k(x)$ having all its singular values more than ϵ and having more than w_k columns and rows. This will contradict the definition of the ϵ numerical rank of a matrix. Thus this particular procedure makes the elements of j th column of the order of ϵ as compared with $\delta(x)$. We can have $r_k - w_k$ matrices of these form making one column of $\hat{A}_k(x)g_i(x)$ small at a time. It is clear that the matrix $\tilde{G}_k(x)$ will thus be nonsingular, square, and an analytic function of x .

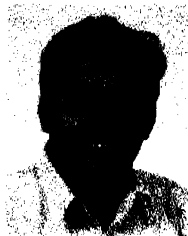
Thus $G_k(x) := \hat{G}_k(x)\tilde{G}_k(x)P_k(x)$ is a square invertible matrix of analytic functions of x .

ACKNOWLEDGMENT

The authors would like to acknowledge helpful discussions with M. Di Benedetto, J. Grizzle, J. Hauser, R. Kadiyala, and P. Kokotović.

REFERENCES

- [1] S. N. Singh, "Control of nearly singular decoupling systems and nonlinear aircraft maneuver," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, no. 6, pp. 775-784, 1988.
- [2] J. Hauser, S. Sastry, and G. Meyer, "Nonlinear control design for slightly nonminimum phase systems-applications to V/STOL aircraft," *Automatica*, vol. 28, no. 4, pp. 665-679, 1992.
- [3] J. J. Romano and S. N. Singh, "I-O map inversion, zero dynamics and flight control," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 26, no. 6, pp. 1022-1029, 1990.
- [4] A. Isidori, *Nonlinear Control Systems*. Springer-Verlag, 2nd ed., 1989.
- [5] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*. Springer-Verlag, 1990.
- [6] J. Descusse and C. Moog, "Decoupling with dynamic compensation for strong invertible affine non-linear systems," *Int. J. Contr.*, vol. 42, no. 6, pp. 1387-1398, 1985.
- [7] S. N. Singh, "Decoupling of invertible nonlinear systems with state feedback and precompensation," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 1237-1239, 1980.
- [8] H. Nijmeijer and J. M. Schumacher, "The regular local non-interacting control problem for nonlinear control systems," *SIAM J. Contr. Optim.*, vol. 24, pp. 1232-1245, 1986.
- [9] H. Nijmeijer and W. Respondek, "Decoupling via dynamic compensation for nonlinear control systems," in *IEEE Contr. Dec. Conf.*, pp. 192-197, 1986.
- [10] J. P. Barbot, N. Pantalos, S. Monaco, and D. Normand-Cyrot, "On the control of singularly perturbed nonlinear systems," in *Proc. IFAC Works. Nonl. Contr. Syst. Des., NOLCOS II*, 1992.
- [11] J. W. Grizzle and M. D. D. Benedetto, "Approximation by regular input-output maps," *IEEE Trans. Automat. Contr.*, vol. 37, no. 7, pp. 1052-1055, 1992.
- [12] S. Sastry, J. Hauser, and P. Kokotović, "Zero dynamics of regularly perturbed systems may be singularly perturbed," *Syst. Contr. Lett.*, vol. 13, pp. 299-314, 1989.
- [13] A. Isidori, S. Sastry, P. Kokotović, and C. Byrnes, "Singularly perturbed zero dynamics of nonlinear systems," *IEEE Trans. Automat. Contr.*, vol. AC-37, no. 10, pp. 1625-1631, 1992.
- [14] J. Descusse and C. Moog, "Dynamic decoupling for right invertible nonlinear systems," *Syst. Contr. Lett.*, vol. 8, pp. 345-349, 1987.
- [15] S. N. Singh, "A modified algorithm for invertibility in nonlinear systems," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 595-598, 1981.
- [16] M. D. Benedetto and J. Grizzle, "Intrinsic notions of regularity for local inversion, output nulling and dynamic extension of nonsquare systems," *Contr. Theory Advan. Tech.*, vol. 6, pp. 357-381, 1990.
- [17] A. Isidori and C. Moog, "On the nonlinear equivalent of the notion of transmission zeros," in *Lecture Notes in Control and Information Sciences*, C. Byrnes and A. Kurzhanski, Eds. New York: Springer-Verlag, 1988, pp. 147-158.
- [18] P. Kokotović, H. Khalil, and J. O'Reilly, *Singular Perturbation Method in Control: Analysis and Design*. Academic Press, 1978.
- [19] J. Huang and W. J. Rugh, "On a nonlinear multivariable servomechanism problem," *Automatica*, vol. 26, no. 6, pp. 963-972, 1990.
- [20] A. Isidori and C. I. Byrnes, "Output regulation of nonlinear systems," *IEEE Trans. Automat. Contr.*, vol. 35, no. 2, pp. 131-140, 1990.
- [21] J. Grizzle, M. D. D. Benedetto, and F. Lamnabhi-Lagarigue, "Necessary conditions for asymptotic tracking in nonlinear systems," University of Michigan, Ann Arbor, Tech. Rep., Control Group report no. GCR 91-4, 1991.
- [22] M. Vidyasagar, *Nonlinear Systems Analysis*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1992.
- [23] R. R. Kadiyala, "AP_LIN: A CAD tool box for nonlinear control design," Ph.D. dissertation, Dept. Elect. Eng., Univ. of California, Berkeley, California, 1992.



Datta N. Godbole (S'93) received the B.E. degree in Electrical Engineering from University of Pune, India, in 1987, and the M. Tech. degree in Systems and Control Engineering from the Indian Institute of Technology, Bombay, India, in 1989. He is currently a candidate for the Ph.D. degree in the Electrical Engineering and Computer Science Department at the University of California, Berkeley.

His research interests include nonlinear control theory, automated highway systems, control of hybrid dynamical systems, applications to transportation systems and vehicle dynamics.

Mr. Godbole is a recipient of the University Gold Medal in engineering from University of Pune, India.



S. Shankar Sastry (S'79-SM'90) received the B.Tech. degree from the Indian Institute of Technology India, in 1977, the M.S. and Ph.D. degrees in electrical engineering in 1979 and 1981, respectively and the M.A. degree in mathematics in 1980, all from the University of California, Berkeley.

He is a Professor in the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, and was a Gordon McKay Professor of Electrical Engineering at Harvard University in 1994. He has held visiting appointments at the Australian National University, Canberra, the University of Rome, the laboratory LAAS in Toulouse, and as a Vinton Hayes Visiting Professor at the Center for Intelligent Control Systems at MIT. His areas of research are nonlinear control, nonholonomic motion planning, robotic remote surgery, control and verification of hybrid systems and biological motor control.

Dr. Sastry is a coauthor (with M. Bodson) of *Adaptive Control: Stability Convergence and Robustness*, (Prentice-Hall, 1989), and (with R. Murray and Z. Li) of *A Mathematical Introduction to Robotic Manipulation*, (CRC Press, 1994). His book, *Nonlinear Control: Analysis, Stability and Control*, is to be published by Addison Wesley in 1995. He is an Associate Editor of the *IMA Journal of Control and Information*, the *International Journal of Adaptive Control and Signal Processing* and the *Journal of Biomimetic Systems and Materials*.

A Generalized Orthonormal Basis for Linear Dynamical Systems

Peter S. C. Heuberger, Paul M. J. Van den Hof, *Member, IEEE*, and Okko H. Bosgra

Abstract—In many areas of signal, system, and control theory, orthogonal functions play an important role in issues of analysis and design. In this paper, it is shown that there exist orthogonal functions that, in a natural way, are generated by stable linear dynamical systems and that compose an orthonormal basis for the signal space ℓ_2^+ . To this end, use is made of balanced realizations of inner transfer functions. The orthogonal functions can be considered as generalizations of, e.g., the pulse functions, Laguerre functions, and Kautz functions, and give rise to an alternative series expansion of rational transfer functions. It is shown how we can exploit these generalized basis functions to increase the speed of convergence in a series expansion, i.e., to obtain a good approximation by retaining only a finite number of expansion coefficients. Consequences for identification of expansion coefficients are analyzed, and a bound is formulated on the error that is made when approximating a system by a finite number of expansion coefficients.

1. INTRODUCTION

CONSIDER a linear time-invariant stable discrete-time system G , represented by its proper transfer function $G(z)$ in the Hilbert space \mathcal{H}_2 , i.e., $G(z)$ is analytic outside the unit circle, $|z| \geq 1$. A general and common representation of $G(z)$ is in terms of its Laurent expansion around $z = \infty$, as

$$G(z) = \sum_{k=0}^{\infty} G_k z^{-k} \quad (1)$$

with $\{G_k\}_{k=0,1,\dots}$ the sequence of Markov parameters.

In constructing this series expansion we have employed a set of orthogonal functions: $\{z^0, z^{-1}, z^{-2}, \dots\}$, where orthogonality is considered in terms of the inner product in \mathcal{H}_2 . In a generalized form we can write (1) as

$$G(z) = \sum_{k=0}^{\infty} L_k f_k(z) \quad (2)$$

with $\{f_k(z)\}_{k=0,1,2,\dots}$ a sequence of orthogonal functions.

There are a number of research areas that deal with the question of either approximating a given system G with a finite number of coefficients in a series expansion as in (2), or (approximately) identifying an unknown system in terms of a finite number of expansion coefficients through

$$\hat{G}(z) = \sum_{k=0}^N L_k f_k(z). \quad (3)$$

The problem that will be analyzed in this paper is the following.

Can we construct a sequence of orthogonal basis functions $\{f_{k,G}(z)\}_{k=0,\dots,\infty}$ with $G \in \mathcal{H}_2$, such that

- to some extent, the basis can be adapted to a linear stable system G to be described, implying that G can be accurately described by only a small number of coefficients in the expansion, and
- the basis allows the construction of an error bound for the approximation of a linear stable system G by a finite length expansion in the basis $f_{k,G}$, i.e., an upper bound on $\|G(z) - \sum_{k=0}^N L_k f_{k,G}(z)\|$ in some prechosen norm, whenever G and \hat{G} do not match exactly.

The use of orthogonal functions with the aim of adapting the system and signal representation to the specific properties of the systems and signals at hand has a long history. The classical work of Lee and Wiener during the 1930's on network synthesis in terms of Laguerre functions [24], [46] is summarized in [25]. Laguerre functions have been used in the 1950's and 1960's to represent transient signals [45], [7]. During the past decades, the use of orthogonal functions has been studied in problems of filter synthesis [22], [30] and for system identification [23], [32], [31], [6] and approximation [35], [36]. In these approaches to system identification, the input and output signals are transformed to a (Laguerre) transformed domain and standard identification techniques are applied to the signals in this domain. Data reduction has been the main motivation in these studies. Identification of continuous-time models with the aid of orthogonal functions is considered in e.g., [38] and [29]. In recent years, a renewed interest in Laguerre functions has emerged. The approximation of (infinite dimensional) systems in terms of Laguerre functions has been considered in [27], [28], [12], [13], and [15]. In the identification of coefficients in finite length series expansions, Laguerre function representations have been considered from a statistical analysis point of view in [43], [42], and [16].

Manuscript received November 17, 1992, revised December 3, 1993. Recommended by Past Associate Editor, B. Pasik-Duncan. This work was supported in part by Shell Research B.V., The Hague, and the Center for Industrial Control Science, The University of Newcastle, Newcastle NSW, Australia.

P. S. C. Heuberger was with the Mechanical Engineering Systems and Control Group, Delft University of Technology, 2 628 CD Delft, The Netherlands and is now with the Dutch National Institute of Public Health and Environmental Protection (RIVM), P.O. Box 1, 3720 BA Bilthoven, The Netherlands.

P. M. J. Van den Hof and O. H. Bosgra are with the Mechanical Engineering Systems and Control Group, Delft University of Technology, Mekelweg 2, 2 628 CD Delft, The Netherlands.

IEEE Log Number 9408269.

The use of Laguerre-function-based identification in adaptive control and controller tuning is studied in [47] and [9]. A second-order extension to the basic Laguerre functions using the so called Kautz functions [21] is subject of discussion in [41] and [44].

In this paper we will expand and generalize the orthogonal functions as basis functions for dynamical system representations. Specifically we will generalize the Laguerre functions and Kautz functions to a situation where a higher degree of flexibility is present in the choice of basis functions, and where consequently a smaller error bound as meant in part b) of the problem can be obtained. Laguerre functions are specifically appropriate for accurate modeling of systems with dominant first-order dynamics, whereas Kautz functions are directed toward systems with dominant second-order resonant dynamics. The generalized basis functions, introduced in this paper, will be suited also for systems with a wide range of dominant dynamics, i.e., dominant high frequency and low frequency behavior.

We will restrict attention to the transfer function space \mathcal{H}_2 being equipped with the usual inner product. This choice, rather than the \mathcal{H}_∞ -space where orthogonality is abandoned, is motivated by the fact that our main intended application of these results is in the area of approximate system identification. As the main stream of approaches in system identification is directed toward prediction error methods and the use of least-squares types of identification criteria, [26], the choice of a two-norm is quite straightforward and natural in this respect.

Note that the two problems a) and b) should be treated as a joint problem. One of the (trivial) solutions to problem a) only is the use of a Gram-Schmidt orthogonalization procedure on the impulse response of the system G itself [1]. In that case the system can be described by a series expansion of only one single term. In this situation, however, no results are available for part b) of the problem.

In an identification context, the use of the orthogonal functions as in (1) leads to the so-called finite impulse response (FIR)-model [26]

$$y(t) = \sum_{k=0}^N G_k(\theta) u(t-k) + \varepsilon(t) \quad (4)$$

where $\varepsilon(t)$ is the one-step-ahead prediction error, and $\{y(t), u(t)\}$ are samples of the output, input of the dynamical system to be identified. The identification of the unknown coefficients $\{G_k(\theta)\}_{k=0, \dots, N}$ through least squares minimization of $\varepsilon(t)$ over the time interval is an identification method that has some favorable properties. First, it is a linear regression scheme, which leads to a simple analytical solution; second, it is of the type of output-error-method, which has the advantage that the input/output system $G(z)$ can be estimated consistently whenever the unknown noise disturbance on the output data is uncorrelated with the input signal [26].

It is well known, however, that for moderately damped systems, and/or in situations of high sampling rates, it may take a large value of N , the number coefficients to be estimated, to capture the essential dynamics of the system G into its model.

If we would be able to improve the basis functions in such a way that an accurate description of the model to be estimated can be achieved by a small number of coefficients in a series expansion, then this is beneficial from both aspects of bias and variance of the model estimate.

For the series expansion in (1) with $f_k = z^{-k}$, it is straightforward to show that a system G will have a finite length series expansion if and only if all system poles are at $z = 0$. Moreover, in the scalar case the length of the expansion, i.e., the index of the last nonzero coefficient, equals the total number of poles at $z = 0$.

As a generalized situation, we can consider Laguerre polynomials [37] that are known to generate a sequence of orthogonal functions [14]

$$f_k(z) = \sqrt{1-a^2} z \frac{(1-az)^k}{(z-a)^{k+1}}, \quad |a| < 1. \quad (5)$$

Similar to above, a system G will have a finite length series expansion if and only if all system poles are at $z = a$, with the length of the expansion being equal to the total number of poles at $z = a$.

In dealing with the problem of finding similar results for any general stable dynamical system $G(z)$, we have considered the question of whether a linear system in a natural way gives rise to a set of orthogonal functions. The answer to this question appears to be affirmative. It will be shown that every stable system gives rise to a complete set of orthonormal functions based on input (or output) balanced realizations, or equivalently based on a singular value decomposition of a corresponding Hankel matrix. These generalized orthogonal basis functions will be shown to provide solutions to problems a) and b).

In Section III we will first briefly state the main result of this paper. Next in Section IV it will be shown how inner functions generate two sets of orthonormal functions that are complete in the signal space ℓ_2 . This is the basic ingredient of the main result. Next an interpretation of these results is given in terms of balanced state-space representations. After showing the relations of the new basis functions with existing ones, we will focus on the dynamics that implicitly are involved in the inner functions generating the basis. It will be shown that if the dynamics of a stable system match the dynamics of the inner function that generates the basis, then the representation of this system in terms of this basis becomes extremely simple. Consequences for a related identification and approximation problem are discussed in Section VIII.

Due to space limitations, a complete statistical analysis of the related system identification problems that result from these basis functions can not be given in this paper. A statistical analysis along similar lines as [43] and [44] is presented elsewhere [39].

The proofs of all results are collected in an appendix.

II. PRELIMINARIES

We will use the following notation.

$(\cdot)^T$ Transpose of a matrix.

$(\cdot)^*$	Complex conjugate transpose of a matrix.
$\mathcal{C}^{p \times m}$	Set of complex-valued matrices of dimension $p \times m$.
$\mathbf{R}^{p \times m}$	Real-valued matrix with dimension $p \times m$.
\mathbb{Z}_+	Set of nonnegative integers.
$\ell_2[0, \infty)$	Space of squared summable sequences on the time interval \mathbb{Z}_+ .
$\ell_2^{m \times n}[0, \infty)$	Space of matrix sequences $\{F_k \in \mathcal{C}^{m \times n}\}_{k=0,1,2,\dots}$ such that $\sum_{k=0}^{\infty} \text{tr}(F_k^* F_k)$ is finite.
$\mathcal{H}_2^{p \times m}$	Set of real $p \times m$ matrix functions, analytic for $ z \geq 1$, that are squared integrable on the unit circle.
$\mathcal{RH}_2^{p \times m}$	Set of real rational $p \times m$ matrix functions, analytic for $ z \geq 1$, that are squared integrable on the unit circle.
$\ \cdot\ _2$	Induced 2-norm or spectral norm of a constant matrix, i.e., its maximum singular value.
$\ \cdot\ _{\infty}$	H_{∞} -norm.
$\text{Vec}(\cdot)$	Vector-operation on a matrix, stacking its columns on top of each other.
(\cdot)	Kronecker matrix product.
$\mathcal{H}(G)$	(Block) Hankel matrix related to transfer function $G = \sum_{k=0}^{\infty} G_k z^{-k}$, defined by $\mathcal{H}_{ij}(G) = G_{i+j-1}$ being the (i, j) -block element.
e_i	i th Euclidian basis vector in \mathbf{R}^n .
I_n	$n \times n$ Identity matrix.

In this paper we will consider discrete-time signals and systems. A linear time-invariant finite-dimensional system will be represented by its rational transfer function $G \in \mathcal{RH}_2^{p \times m}$, with m the number of inputs in u , and p the number of outputs in y . State-space realizations will be considered of the form

$$x(k+1) = Ax(k) + Bu(k) \quad (6)$$

$$y(k) = Cx(k) + Du(k) \quad (7)$$

with $A \in \mathcal{C}^{n \times n}$, $B \in \mathcal{C}^{n \times m}$, $C \in \mathcal{C}^{p \times n}$, and $D \in \mathcal{C}^{p \times m}$. (A, B, C, D) is an n -dimensional realization of G if $G(z) = C(zI - A)^{-1}B + D$. A realization is stable if all eigenvalues of A lie strictly within the unit circle. If a realization is stable, the controllability gramian P and observability gramian Q are defined as the solutions to the Lyapunov equations $PA^* + BB^* = P$ and $A^*QA + C^*C = Q$, respectively. A stable realization is called (internally) balanced if $P = Q = \Sigma$, with $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$, $\sigma_1 \geq \dots \geq \sigma_n$, a diagonal matrix with the positive Hankel singular values as diagonal elements. A stable realization is called input balanced if $P = I$, $Q = \Sigma^2$, and output balanced if $P = \Sigma^2$, $Q = I$.

A system $G \in \mathcal{RH}_2^{p \times m}$ is called inner if it satisfies $G^*(z^{-1})G(z) = I$. As G is analytic outside and on the unit circle, it has a Laurent series expansion $\sum_{k=0}^{\infty} G_k z^{-k}$.

III. THE MAIN RESULT

We will start the technical part of this paper by giving the basic result first and then consecutively give the analysis that provides the ingredients for making the result plausible.

Theorem 3.1: Let G be an $m \times m$ inner transfer function with McMillan degree $n > 0$, having a Laurent expansion $G(z) = \sum_{k=0}^{\infty} G_k z^{-k}$ and satisfying $\|G_0\|_2 < 1$, and let (A, B, C, D) be a balanced realization of $G(z)$. Denote

$$V_k(z) = (zI - A)^{-1}BG^k(z). \quad (8)$$

Then the set of functions $\{V_k(z)\}_{k=1, \dots, n, k=0, \dots, \infty}$ constitutes an orthonormal basis of the function space $\mathcal{H}_2^{1 \times m}$.

A direct consequence of this theorem is the following corollary. \square

Corollary 3.2: Let G be an inner function with McMillan degree n as in Theorem 3.1, with a corresponding sequence of basis functions $V_k(z)$. Then for every proper stable transfer function $H \in \mathcal{H}_2^{p \times m}$ there exist unique $D_q \in \mathbf{R}^{p \times m}$, and $L = \{L_k\}_{k=0,1,\dots} \in \ell_2^{p \times n}[0, \infty)$, such that

$$H(z) = D_q + z^{-1} \sum_{k=0}^{\infty} L_k V_k(z). \quad (9)$$

We refer to D_q, L_k as the orthogonal expansion coefficients of $H(z)$. \square

Note that due to the fact that $V_k(z)$ is an $n \times m$ -matrix of transfer functions, the dimension of each L_k is $p \times n$.

IV. ORTHONORMAL FUNCTIONS GENERATED BY INNER TRANSFER FUNCTIONS

In this section we will show that a square and inner transfer function gives rise to an infinite set of orthonormal functions. This derivation is based on the fact that a singular value decomposition of the Hankel matrix associated to a linear system induces a set of left (right) singular vectors that are orthogonal. Considering the left (right) singular vectors as discrete time functions, they are known to be orthogonal in ℓ_2 -sense, thus generating a number of orthogonal functions being equal to the McMillan degree of the corresponding system. We will embed an inner function with McMillan degree n into a sequence of inner functions with McMillan degree kn , for which the left (right) singular vectors of the Hankel matrix span a space with dimension kn . If we let $k \rightarrow \infty$ the set of left (right) singular vectors will yield an infinite number of orthonormal functions, which can be shown to be complete in ℓ_2 .

First we have to recapitulate some properties of inner transfer functions.

Proposition 4.1: Let $G(z)$ be an inner transfer function with a Laurent expansion $G(z) = \sum_{k=0}^{\infty} G_k z^{-k}$. Then

$$\sum_{k=0}^{\infty} G_{k+i}^T G_k = I \quad \text{for } i = 0; \quad (10)$$

$$= 0 \quad \text{for } i > 0. \quad (11)$$

\square

The Hankel matrix of an inner transfer function has some specific properties, reflected in the following two results.

Proposition 4.2: Let $G(z)$ be an inner function with McMillan degree $n > 0$. Then a singular value decomposition (svd) of $\mathcal{H}(G)$ satisfies

$$\mathcal{H}(G) = U_0 V_0^*$$

with $U_0, V_0 \in \mathcal{C}^{\infty \times n}$ unitary,¹ and the pair (U_0, V_0) is unique modulo postmultiplication with a unitary matrix $T \in \mathcal{C}^{n \times n}$. \square

The proposition states that an inner transfer function has all Hankel singular values equal to one. For continuous-time systems, this is proven in [11]. The discrete-time version follows straightforwardly by applying a bilinear transformation.

Proposition 4.3: Let $G(z)$ be a square inner function, having a Laurent expansion $G(z) = \sum_{k=0}^{\infty} G_k z^{-k}$. Denote the block Toeplitz matrices

$$T_v = \begin{bmatrix} G_0 & G_1 & G_2 & \cdots & \cdots \\ 0 & G_0 & G_1 & G_2 & \cdots \\ 0 & 0 & G_0 & G_1 & \cdots \\ \vdots & \vdots & 0 & G_0 & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad (12)$$

$$T_u = \begin{bmatrix} G_0 & 0 & 0 & \cdots & \cdots \\ G_1 & G_0 & 0 & \cdots & \cdots \\ G_2 & G_1 & G_0 & \ddots & \cdots \\ \vdots & \vdots & G_1 & G_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}. \quad (13)$$

Then

- i) $T_v T_v^T = T_u^T T_u = I$;
- ii) $T_v V_0 = T_u^T U_0 = 0$, for any unitary matrices U_0, V_0 satisfying $U_0 V_0^* = \mathcal{H}(G)$. \square

Lemma 4.4: Let $G(z)$ be a square inner function with McMillan degree n . Then for all $k \in \mathbb{Z}_+$, $G^k(z)$ is an inner function with McMillan degree kn . \square

The result in the lemma is quite straightforward if one realizes that an inner function has all poles within the unit circle, and all zeros outside the circle.

Considering Proposition 4.2, it follows that the rows of V_0^* and the columns of U_0 , are n mutually orthonormal vectors of infinite dimension. Additionally Lemma 4.4 shows that we can construct an inner transfer function with increasing McMillan degree, by repeatedly multiplying the transfer function with itself, and thus implicitly creating an increasing number of orthogonal vectors. The following result shows how we can increase this number of vectors, by embedding the svd of $\mathcal{H}(G)$ into a sequence of svd's of $\mathcal{H}(G^k)$.

Theorem 4.5: Let $G(z)$ be a square inner function with McMillan degree $n > 0$. Then

- a) There exist unitary matrices $U_i, V_i \in \mathcal{C}^{\infty \times n}$, $i = 0, 1, \dots$, such that for every $0 \neq k \in \mathbb{Z}_+$, the matrices

$$\Gamma_k^o = [U_{k-1} \cdots U_1 U_0] \quad \text{and} \quad (14)$$

$$\Gamma_k^c = \begin{bmatrix} V_0^* \\ V_1^* \\ \vdots \\ V_{k-1}^* \end{bmatrix} \quad (15)$$

¹With slight abuse of notation we will use this notation to indicate an operator $\mathcal{C}^n \rightarrow \ell_2[0, \infty)$.

constitute a singular value decomposition of $\mathcal{H}(G^k)$, through

$$\mathcal{H}(G^k) = \Gamma_k^o \Gamma_k^c. \quad (16)$$

- b) The matrix sequence $\{U_i, V_i\}_{i=0,1,\dots}$ is unique up to postmultiplication of each U_i and V_i with one and the same unitary matrix.
- c) Let $G(z)$ have a Laurent expansion $G(z) = \sum_{i=0}^{\infty} G_i z^{-i}$, and consider the block Toeplitz matrices T_u, T_v as in (12), (13) then the matrix sequence $\{U_i, V_i\}_{i=0,1,\dots}$ satisfies

$$V_k^* = V_{k-1}^* T_v \quad (17)$$

$$U_k = T_u U_{k-1} \quad \text{for } k = 1, 2, \dots \quad (18)$$

\square

The theorem shows the construction of orthogonal matrices Γ_k^o, Γ_k^c that have a nesting structure. The suggested svd of $\mathcal{H}(G^k)$ incorporates svd's of $\mathcal{H}(G^i)$ for all $i < k$. In this way orthogonal matrices Γ_k^o and Γ_k^c are constructed with an increasing rank. Note that the restriction on the structure of the consecutive svd's is so strong that, according to b), given a singular value decomposition $\mathcal{H}(G) = U_0 V_0^*$, the matrix sequence $\{U_i, V_i, i = 1, 2, \dots\}$ is uniquely determined. Note also that there is a clear duality between the controllability part Γ_k^c and the observability part Γ_k^o . To keep the exposition and the notation as simple as possible we will further restrict attention to the controllability part of the problem. Dual results exist for the observability part.

Proposition 4.6: Let $G(z)$ be an $m \times m$ inner function with McMillan degree $n > 0$, and consider any sequence of unitary matrices $\{V_i\}_{i=0,1,\dots}$ satisfying (17) in Theorem 4.5. Denote for $k \in \mathbb{Z}_+$

$$V_k(z) = \sum_{i=0}^{\infty} M_k(i) z^{-i}, \quad \text{with } M_k(i) \in \mathcal{C}^{m \times m} \text{ defined by} \quad (19)$$

$$V_k^* = [M_k(0) \ M_k(1) \ M_k(2) \ \cdots].$$

Then

$$V_k(z) = V_0(z) G^k(z).$$

\square

The proposition actually is a z -transform-equivalent of the result in Theorem 4.5. It shows the construction of the controllability matrix Γ_k^c .

In the next stage we show that this controllability matrix generates a sequence of orthogonal functions that is complete in ℓ_2^n .

Theorem 4.7: Let $G(z)$ be an $m \times m$ inner function with McMillan degree $n > 0$, such that $\|G_0\|_2 < 1$; consider a sequence of unitary matrices $\{V_i\}_{i=0,1,\dots}$ as meant in Theorem 4.5. For each $k \in \mathbb{Z}_+$ consider the function $\phi_k : \mathbb{Z}_+ \rightarrow \mathcal{C}^m$ defined by

$$[\phi_k(0) \ \phi_k(1) \ \phi_k(2) \ \cdots] = V_k^*.$$

Then the set of functions $\Psi(G) := \{\phi_k\}_{k=0}^{\infty}$ constitutes an orthonormal basis of the signal space $\ell_2^n[0, \infty)$. \square

Proposition 5.5: Let $G(z)$ be an $m \times m$ inner transfer function with McMillan degree $n > 0$, whose Hankel matrix has an svd $\mathcal{H}(G) = U_0 V_0^*$; let (A, B, C, D) be a minimal balanced realization of G such that $V_0^* = [B \ AB \ A^2 B \ \dots]$. Then the unique sequence of orthogonal matrices $\{\Gamma_k^c\}_{k=1,2,\dots}$ as considered in Theorem 4.5 is determined by

$$\Gamma_k^c = [B_k \ A_k B_k \ A_k^2 B_k \ \dots] \quad (26)$$

with A_k, B_k as defined in (22), (23). \square

The above result shows how a minimal balanced realization of G actually generates the sequence of orthogonal matrices Γ_k^c , the rows of which are the basis functions in our orthonormal basis of ℓ_2^n .

We will show that there exist recursive formulae for constructing the orthogonal functions.

Proposition 5.6: Let G be an inner function, $G \in \mathcal{G}_1$, and consider the assumptions and notation as in Theorem 4.5 and Proposition 5.5. Denote

$$X = BC^* \quad \text{and} \quad (27)$$

P any matrix satisfying

$$PB = BD. \quad (28)$$

Then the elements of Γ_k^c are determined by the following recursive equations

$$M_0(0) = B \quad (29)$$

$$M_k(i+1) = AM_k(i) + \sum_{j=1}^k P^{j-1} X M_{k-j}(i), \quad i \geq 0; \quad (30)$$

$$M_k(0) = PM_{k-1}(0) \quad (31)$$

with Γ_k^c as in (15) with

$$V_k^* = [M_k(0) \ M_k(1) \ M_k(2) \ \dots]$$

as in (19). \square

The recursive equations show how we can simply construct the set of orthogonal functions. Note that the matrix P in (28) is nonunique. The result (29)-(31) however is unique. A straightforward choice for P satisfying (28) is

$$P = BD(B^*B)^{-1}B^*. \quad (32)$$

Note that, as a result of Proposition 5.3, the matrix B^*B is invertible whenever $G \in \mathcal{G}_1$.

The orthogonal functions $\Psi(G)$ generated by an inner function G can be represented in terms of their generating functions $V_k(z)$, as defined in Proposition 4.6. These generating transfer functions can also be realized in terms of a minimal balanced realization of G . This is reflected in the following theorem.

Theorem 5.7: Let G be an inner function, $G \in \mathcal{G}_1$, with a minimal balanced realization (A, B, C, D) . Let this inner function generate an orthonormal basis with corresponding generating functions $V_k(z)$, as defined in Proposition 4.6.

1) Let F be a matrix determined by

$$F = X - PA \quad (33)$$

with X defined in (27) and P any matrix satisfying (28). Then, for $k \in \mathbb{Z}_+$,

- a) $V_k(z) = [(zI - A)^{-1}F(I - zA^*)]^k z(zI - A)^{-1}I$
- b) $V_k(z)$ is unique, i.e., it is not dependent on the specific choice of P in (28).

2) If there exists a matrix R such that $B = RC^*$, then $F = R$ satisfies the conditions of Part 1 of this theorem \square

Now we come to the construction of a series expansion of any stable proper rational transfer function, in terms of the new orthonormal basis.

Theorem 5.8: Let G be an inner function, $G \in \mathcal{G}_1$, with a minimal balanced realization (A, B, C, D) . Let this inner function generate an orthonormal basis with corresponding generating functions $V_k(z)$, as defined in Proposition 4.6. Let $H \in \mathcal{H}_2^{p \times m}$ be any proper and stable transfer function with a minimal realization (A_s, B_s, C_s, D_s) . Then

$$H(z) = D_s + z^{-1} \sum_{k=0}^{\infty} L_k V_k(z) \quad (34)$$

with $L_k \in \mathbb{C}^{p \times n}$ determined by

$$L_k = C_s Q_k \quad (35)$$

$$Q_0 = A_s Q_0 A_s^* + B_s B_s^* \quad (36)$$

$$Q_{i+1} = A_s Q_{i+1} A_s^* + A_s Q_i F^* - Q_i A F^* \quad (37)$$

with F as defined in (33). \square

In Section VII we will show that specific choices of $G(z)$ in relation with $H(z)$, i.e., specific relations between the inner function G producing the orthonormal basis and a transfer function H that should be described in this basis, will lead to very simple representations.

VI. A GENERALIZATION OF CLASSICAL BASIS FUNCTIONS

In this section we show three examples of well-known sets of orthogonal functions that are frequently used in the description of linear time-invariant dynamical systems and that occur as special cases in the framework that is discussed in this paper.

Pulse Functions

Consider the inner function $G(z) = z^{-1}$, $G \in \mathcal{G}_1$. The Hankel matrix of G satisfies

$$\begin{aligned} \mathcal{H}(G) &= \begin{bmatrix} 1 & 0 & \dots & \cdot & 0 \\ 0 & 0 & 0 & \cdot & 0 \\ 0 & 0 & 0 & \cdot & 0 \\ \vdots & \vdots & \cdot & \ddots & \cdot \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix} [1 \ 0 \ 0 \ \dots] = U_0 V_0^*. \end{aligned} \quad (38)$$

As a result $V_0(z) = 1$, and with Proposition 4.6 the generating transfer functions $V_k(z)$ satisfy $V_k(z) = G^k(z) = z^{-k}$, $k = 0, 1, \dots$. The corresponding set of basis functions $\Psi(G)$

determined by $\phi_k(t) = \delta(t - k)$ with $\delta(\tau)$ the Kronecker delta function.

The inner function G can be realized by the minimal balanced realization $(A, B, C, D) = (0, 1, 1, 0)$. The equation $PB = BD$ is satisfied by $P = 0$, and the corresponding result for F is $F = BC = 1$. Applying Theorem 5.8 shows the classical result that $L_k = C_s A_s^k B_s$.

Laguerre Functions

Consider the inner function $G(z) = \frac{1 - az}{z - a}$, with some real-valued a , $|a| < 1$, and denote $\eta = 1 - a^2$. A minimal balanced realization of G is given by $(A, B, C, D) = (a, \sqrt{\eta}, \sqrt{\eta}, -a)$. Equation $PB = BD$ is satisfied by $P = -a$, leading to $F = BC - PA = \eta + a^2 = 1$. Taking account of the fact that for one-dimensional scalar G , $M_k(i) = \phi_k(i)$, it follows from Proposition 5.6 that

$$\phi_0(0) = \sqrt{\eta} \quad (39)$$

$$\phi_k(i+1) = a\phi_k(i) + \eta \sum_{j=1}^k (-a)^{j-1} \phi_{k-j}(i) \quad (40)$$

$$\phi_k(0) = -a\phi_{k-1}(0). \quad (41)$$

These equations exactly match the equations that generate the normalized discrete-time Laguerre polynomials with discount factor a , [14], [32].

The corresponding generating transfer functions $V_k(z)$ can be analyzed with the result of either Proposition 4.6 or Theorem 5.7

$$V_k(z) = \sqrt{\eta} z \frac{(1 - az)^k}{(z - a)^{k+1}}. \quad (42)$$

This exactly fits with the formulation of the generating transfer functions of discrete-time Laguerre polynomials in, e.g., [23].

Kautz Functions

Consider the inner function $G(z) = \frac{-cz^2 + b(c-1)z + 1}{z^2 + b(c-1)z - c}$ with some real-valued b, c satisfying $|c|, |b| < 1$.

A balanced realization of $G(z)$ can be found to be equal to

$$A = \begin{bmatrix} b & \sqrt{(1-b^2)} \\ c\sqrt{(1-b^2)} & -bc \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \sqrt{(1-c^2)} \end{bmatrix}$$

$$C = [\gamma_2 \quad \gamma_1] \quad D = -c$$

with $\gamma_1 = -b\sqrt{(1-c^2)}$ and $\gamma_2 = \sqrt{(1-c^2)(1-b^2)}$.

With the expression for $V_0(z)$ from Theorem 5.7-a) it follows that

$$z^{-1}V_0(z) = \frac{\sqrt{(1-c^2)}}{z^2 + b(c-1)z - c} \begin{bmatrix} \sqrt{(1-b^2)} \\ z - b \end{bmatrix} \quad (43)$$

which exactly equals $\begin{bmatrix} \psi_2(z, b, c) \\ \psi_1(z, b, c) \end{bmatrix}$ representing the orthogonal Kautz functions, as represented in [41], [44]. Postmultiplication with $G^k(z)$ is equivalent to the situation in the case of Kautz functions.

VII. ORTHONORMAL FUNCTIONS ORIGINATING FROM GENERAL DYNAMICAL SYSTEMS

We have shown that any square inner transfer function $G \in \mathcal{G}_1$ generates an orthonormal basis for the signal space ℓ_2 . One of the reasons for developing this generalized bases was to find out whether we can yield a more suitable representation of a general dynamical system, when the basis within which we describe the system is more or less adapted to the system dynamics. In view of the results presented so far, this aspect relates to the question whether we can construct an inner transfer function generating a basis that incorporates dynamics of a general system to be represented within this basis.

There are several ways of connecting general transfer functions to inner functions, as e.g., inner/outer factorization [10], [5], normalized coprime factorization [8], [40], [33], [4], or inner-unstable factorization [2]. Even if the corresponding inner functions are not square, they can always be embedded in a square inner function [11]. In this paper, however, we will explore a different connection, where a general stable dynamical system with input balanced realization (A, B, C, D) will induce a square inner function through retaining the matrices (A, B) and constructing (C', D) such that (A, B, C', D) is inner. This implies that the poles of the stable dynamical system are retained in the corresponding inner function. The following result shows the existence and construction of such an inner function.

Proposition 7.1: Let (A, B) be the system matrix and input matrix of an input balanced realization of a transfer function $H \in \mathcal{RH}_2^{p \times m}$ with McMillan degree $n > 0$, and with rank $B = m$. Then

a) There exist matrices C', D such that (A, B, C', D) is a minimal balanced realization of a square inner function $G' \in \mathcal{G}_1$.

b) A realization (A, B, C', D) has the property mentioned in a) if and only if

$$C' = UB^*(I_n + A^*)^{-1}(I_n + A) \quad (44)$$

$$D = U[B^*(I_n + A^*)^{-1}B - I_m] \quad (45)$$

with $U \in \mathbb{R}^{m \times m}$ any unitary matrix.

c) For a realization satisfying (44), (45) a valid choice of matrix F satisfying (33) is given by

$$F = [I_n + B(U - I_m)(B^*B)^{-1}B^*](I_n + A)(I_n + A^*)^{-1}. \quad (46)$$

□

In the proposition all inner functions are characterized that can be constructed in the way as described above, by retaining the matrices (A, B) of any given stable system. Note that the extension C', D is not unique. The nonuniqueness is reflected by a possible unitary premultiplication of the inner function. Note also that when choosing $U = I_m$, expression (46) reduces to $F = (I_n + A)(I_n + A^*)^{-1}$.

We will now present a result that is very appealing. It shows that when we want to describe the dynamical system H in terms of the basis that it has generated, as presented in Proposition 7.1, then the series expansion in the new orthogonal basis becomes extremely simple.

Theorem 7.2: Let $H \in \mathcal{RH}_2^{p \times m}$ have an input balanced realization (A_s, B_s, C_s, D_s) , having all controllability indexes > 0 . Let $G \in \mathcal{G}_1$ be a square inner function with minimal balanced realization (A, B, C, D) such that $A = A_s$ and $B = B_s$, generating an orthonormal basis with generating transfer functions $V_k(z)$. Then

$$H(z) = D_s + z^{-1} \sum_{k=0}^{\infty} L_k V_k(z) \quad (47)$$

with

$$L_0 = C_s' \quad \text{and} \quad (48)$$

$$L_k = 0 \quad \text{for } k > 0. \quad (49)$$

Proof: The proof follows by applying Theorem 5.8. With $A = A_s$, $B = B_s$ (36) becomes $Q_0 = A Q_0 A^* + B B^*$. Since (A, B) is input balanced, the solution to this equation is $Q_0 = I$, leading to $L_0 = C_s'$. Substituting $Q_0 = I$ in (37) and using the stability of A shows that $Q_i = 0$ for $i > 0$. \square

The theorem shows that when we use a general stable and proper dynamical system to generate an orthonormal basis as described above, then the system itself has a very simple representation in terms of this basis. It is represented in a series expansion with only two nonzero expansion coefficients, being equal to the system matrices C_s' and D_s .

In the next section we will discuss the results of this paper regarding their relevance to problems of system identification and system approximation.

It has to be stressed that, so far, we have only used the generalized orthonormal basis to study the series expansion of a given stable transfer function. Similar to the case of the pulse functions and Laguerre functions, the presented generalized functions induce a transformation of ℓ_2 -signals to a transform domain, compare e.g., with the z -domain when pulse functions are used. In this transform domain dynamical system equations can be derived, leading to transform pairs of time-domain and orthogonal-domain system representations. In the case of a Laguerre basis, these kinds of transformations actually have been used frequently also in an identification context, by first transforming the measured input/output signals to the Laguerre domain, and consecutively identifying a system in this domain; see e.g., [22], [23], [32], [31].

For the generalized basis, results along these lines have been presented in [18], [19]. An analysis of the system transformations between time domain and generalized transform domain is treated in [19] and [39].

VIII. SYSTEM APPROXIMATION AND IDENTIFICATION

We will now discuss the way in which the introduced orthogonal basis functions provide a solution to problem b) as mentioned in the introduction, i.e., the quantification of an error bound for finite length expansion approximants.

We will present results showing that the speed of convergence in an orthogonal series expansion can be quantified and that an increase of speed is obtained as the dynamics of system and basis approach each other. To formulate these results we need an alternative formulation of Theorem 5.8 in terms of Kronecker products.

Proposition 8.1: Let $H \in \mathcal{RH}_2^{p \times m}$ be a transfer function with an input balanced realization (A_s, B_s, C_s, D_s) , and let (A, B) be an input balanced pair that generates an $m \times 1$ inner transfer function $G \in \mathcal{G}_1$, leading to an orthonormal basis $\Psi(G)$.

Then the orthogonal expansion coefficients L_k satisfying $H(z) = D_s + z^{-1} \sum_{k=0}^{\infty} L_k V_k(z)$ are determined by

$$Vec(L_k) = Z X^k Y \quad (50)$$

with

$$Z = (I \otimes C_s) M^{-1} \quad (51)$$

$$Y = Vec(B_s B^*) \quad (52)$$

$$X = N M^{-1} \quad (53)$$

$$M = I \otimes I - A \otimes A_s \quad (54)$$

$$N = F \otimes A_s - F A^* \otimes I. \quad (55)$$

Note that due to (50) we can consider $Vec(L_k)$ as a sequence of Markov parameters of a dynamical system with a state-space realization given by $(X, Y, Z, 0)$. By examining the eigenvalues of this realization, we create the possibility of drawing some conclusions on the speed of convergence of the series expansion. The following result is taken from [19].

Proposition 8.2: Consider the situation of Proposition 8.1 with $H(z)$ and $G(z)$ having McMillan degree n_s , n , respectively, and $m = 1$. Let μ_i , $i = 1, \dots, n_s$ denote the eigenvalues of A_s , and ρ_j , $j = 1, \dots, n$ denote the eigenvalues of A . The dynamical system $Z(zI - X)^{-1}Y$ has a realization $(X_o, Y_o, Z_o, 0)$ that satisfies

a) X_o has dimension n_s ;

b) X_o has eigenvalues λ_i , $i = 1, \dots, n_s$, that satisfy

$$|\lambda_i| = \prod_{j=1}^n \left| \frac{\mu_i - \rho_j}{1 - \mu_i \rho_j} \right| \quad (56)$$

Since the proof of this proposition is somewhat outside the scope of this paper, the reader is referred to [19].

The above proposition shows that we can draw conclusions on the convergence rate of the sequence of expansion coefficients $\{L_k\}_{k=0}^{\infty}$, when given the eigenvalues of the original system $H(z)$ and the eigenvalues of the inner function $G(z)$ that generates the basis. Note that when the sets of eigenvalues $\{\mu_i\}$, $\{\rho_j\}$ coincide, then $\lambda_i = 0$, for all i , and consequently the sequence $\{L_k\}$ will have a finite number of elements unequal to zero. The above result also enables the determination of an upper bound on the error that is made, when we approximate a given system $H(z)$ through a finite number of its expansion coefficients.

Theorem 8.3: Consider the situation of Proposition 8.2, and denote

$$\hat{H}^N(z) = D_s + z^{-1} \sum_{k=0}^{N-1} L_k V_k(z)$$

and $\lambda := \max_i |\lambda_i|$. Then there exists a finite $c \in \mathbb{R}$ such that for any $\eta \in \mathbb{R}$, $\eta > \lambda$

$$\|H(z) - \hat{H}^N(z)\|_{\infty} \leq c \frac{\eta^{N+1}}{1 - \eta}. \quad (57)$$

Since λ is a measure for the "closeness" of system dynamics and basis dynamics, the above theorem shows that the error that is made when neglecting the tail of a series expansion, becomes smaller as λ becomes smaller. As a result, when restricting to a fixed number of expansion coefficients, the approximation error gets smaller the more accurate the basis dynamics is "adapted" to the system.

In the final part of this paper we will briefly comment on how these results could be employed in an approximate identification framework. As mentioned in the introduction, identification of a finite impulse model (FIR) (4), has some important advantages; however, it fails to be successful when the number of coefficients to be estimated becomes large. This may happen in situations of high sampling rates, moderately damped systems, as well as systems that have dominant dynamics in both the high-frequent and low-frequent region (e.g., multitime-scale systems). An alternative way to attain the advantages of this identification method, is to exploit the model structure

$$y(t) = D(\theta) + \sum_{k=0}^{N-1} L_k(\theta) V_k(q) u(t) + \varepsilon(t) \quad (58)$$

where $\varepsilon(t)$ is the one-step-ahead prediction error, $D(\theta)$, $L_k(\theta)$ the parameterized expansion coefficients, and with $V_k(z)$ representing an appropriately chosen basis.

Note that this model structure can simply be written as

$$y(t) = D(\theta) + \sum_{k=0}^{N-1} L_k(\theta) u_k(t) + \varepsilon(t) \quad (59)$$

where $u_k(t)$ can simply be calculated by applying $u(t)$ to the known-filters $V_k(q)$, compare Fig. 1.

Identifying θ through least squares optimization of $\varepsilon(t)$ over the time interval, is a similar problem as in the case of a FIR-model. With appropriately chosen basis functions, however, the convergence rate of the series expansion can become extremely fast; with only a few coefficients to be estimated a very accurate approximate model can be obtained. This is of course interesting and appealing from both aspects of bias (accurate approximation is possible) and variance (few parameters to be estimated from data). An analysis of bias and variance errors in these identification schemes is presented in [39].

Additionally, when comparing these "orthogonal FIR" model structures with nonlinearly parameterized model structure as e.g., a Box Jenkins or ARMAX model ([26]), we avoid problems of possible occurrence of local (nonglobal) minima in the quadratic identification criterion. Moreover, the freedom in the choice of basis functions allows the useful use of "a priori information" concerning the system dynamics.

Very often an identification experimenter has a rough-knowledge about the dynamics of the system under consideration, e.g., from previous experiments or from physical insight on the process dynamics. It would be favorable to exploit

this knowledge in an identification procedure. The method suggested above, shows that this a priori knowledge can be exploited in terms of the basis functions that are chosen. When we have rough-knowledge about the poles of the system, we can construct basis functions that are based on this set of poles. The more accurate the poles are, i.e., the more accurate our a priori information is, the better we can adapt the basis functions to the system dynamics. As a result, see Theorem 8.3, the estimated model can become more accurate when restricting to a prespecified number of coefficients to be estimated.

Effectively the identification problem now reflects the identification of the mismatch between the system under consideration and the knowledge that already was available, represented in the basis functions. This actually is very appealing, as the a priori information simplifies the identification procedure. Note that in the way described above, the *a priori* information does not have to be exact, i.e., it is not of the type of fixing *a priori* a constraint on the model parameters, as e.g., the steady-state gain. The information can be uncertain. The only result is that the more accurate it is, the more simple the system representation will be.

This discussion also motivates the use of an iterative scheme, where the identification of parameters θ is performed iteratively, using the model that is estimated in step $i-1$ for constructing the basis functions for step i . An example of such an iterative scheme has been shown in [19].

One remark that has to be made in this respect, is a remark on the model order of a system represented by a finite number of expansion coefficients. The McMillan degree of this system, as in the case of an FIR-representation, will generally be large. This results from the following observation.

Proposition 8.4 Consider the transfer function

$$\hat{H}^N(z) = \hat{D} + \sum_{k=0}^{N-1} L_k V_k(z)$$

with $V_k(z)$ the generating transfer functions of an orthonormal basis $\Psi(G)$, where the inner function $G \in \mathcal{G}_1$ has a minimal balanced realization (A, B, C, D) with dimension n . Then $\hat{H}^N(z)$ has a state-space realization $(A_{N-1}, B_{N-1}, K, \hat{D})$, with A_{N-1}, B_{N-1} defined in (22), (23), and $K = [\hat{L}_0 \ L_1 \ \hat{L}_2 \ \cdots \ L_{N-1}]$. \square

The proof of this proposition follows by inspection.

With \hat{L}_i being the result of an unconstrained optimization in an identification procedure, the state-space dimension of the model will generically be equal to Nn . Consequently, if one wants to represent the model again in a traditional state-space form of low dimension, a model reduction procedure will have to be used to arrive at a reduced dimension. This also motivates a further analysis of the realization problem in terms of orthogonal expansion coefficients $\{L_k\}$.

IX. CONCLUSIONS

We have developed a theory on orthogonal functions as basis functions for general linear time-invariant stable systems. The basic ingredient is that every square inner transfer function in

a very natural way induces two sets of orthogonal functions that form a basis of the signal space ℓ_2 . The ordinary pulse functions and the classical Laguerre and Kautz polynomials are special cases in this theory of inner functions.

With this concept we have explored the connection between a general dynamical system and an inner function, by letting the inner function be determined through a specified set of poles. An important property of the resulting orthonormal functions is that they—to some extent—incorporate the dynamic behavior of the underlying system. We have developed a theory on these system based orthogonal functions, both on an input-output level and in terms of balanced state-space realizations. Furthermore we have shown how the alternative basis can be fruitfully used in problems of system approximation and identification, leading to simplified identification schemes, in which *a priori* knowledge about the process dynamics can be utilized by incorporating the information into the basis.

APPENDIX

Lemma A1: Let $G(z)$, $F(z)$, and $R(z)$ be stable transfer functions with Laurent expansions $G(z) = \sum_{k=0}^{\infty} G_k z^{-k}$, $F(z) = \sum_{k=0}^{\infty} F_k z^{-k}$, and $R(z) = \sum_{k=0}^{\infty} R_k z^{-k}$. Then $R(z) = F(z)G(z)$ if and only if

$$[R_0 \ R_1 \ R_2 \ \dots] = [F_0 \ F_1 \ F_2 \ \dots] T_v \quad (\text{A.1})$$

with T_v as defined in (12). \square

Proof: The equality $R(z) = (\sum_{k=0}^{\infty} F_k z^{-k}) (\sum_{k=0}^{\infty} G_k z^{-k})$ is equivalent to $\sum_{k=0}^{\infty} R_k z^{-k} = \sum_{k=0}^{\infty} [\sum_{i=0}^k F_i G_{k-i}] z^{-k}$, and to $R_k = \sum_{i=0}^k F_i G_{k-i}$, which exactly matches (A.1). \square

Lemma A2 [11]: Given matrices $X \in \mathbb{C}^{n \times m}$, $Y \in \mathbb{C}^{n \times r}$, $r \geq m$, with $XX^* = YY^*$; then there exists a $W \in \mathbb{C}^{m \times r}$ such that $Y = XW$ and $WW^* = I$. \square

Lemma A3: Let $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Then $\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$ and $\sum_{i=0}^t \binom{k+i}{k} = \binom{k+t+1}{k+1}$. \square

Proof: By simple calculation. \square

Lemma A4: Let $G(z)$ be a square inner function such that $\|G_0\| < 1$, with $\|\cdot\|$ any induced matrix norm. Let G^k have a Laurent expansion $G^k(z) = \sum_{i=0}^{\infty} G_i^{(k)} z^{-i}$. Then $\|G_i^{(k)}\| \leq \binom{k+i-1}{k-1} \|G_0\|^{k-i}$ for $0 \leq i \leq k-1$. \square

Proof: The proof will be given by induction. For $k=1$, validity is trivial. Suppose that the statement holds true for $k \leq n$. Now we consider two cases.

i) Consider $G_i^{(n+1)}$, where $i < n$

$$\begin{aligned} G_i^{(n+1)} &= G_0 G_i^{(n)} + G_1 G_{i-1}^{(n)} + \dots + G_i G_0^{(n)}; \\ \|G_i^{(n+1)}\| &\leq \|G_0\| \|G_i^{(n)}\| + \|G_1 G_{i-1}^{(n)}\| + \dots + \|G_i G_0^{(n)}\| \\ &\leq \|G_0\| \binom{n+i-1}{n-1} \|G_0\|^{n-i} + \\ &\quad + \binom{n+i-2}{n-1} \|G_0\|^{n-i+1} + \\ &\quad + \dots + \binom{n}{n-1} \|G_0\|^{n-1} + \|G_0\|^n \end{aligned} \quad (\text{A.2})$$

$$\begin{aligned} &\leq \|G_0\|^{n+1-i} \sum_{j=0}^i \binom{n-1+j}{n-1} \\ &= \|G_0\|^{n+1-i} \binom{n+i}{n} \quad \text{by Lemma A3.} \end{aligned} \quad (\text{A.3})$$

ii) Consider the case $i = n$

$$G_n^{(n+1)} = G_0 G_n^{(n)} + G_1 G_{n-1}^{(n)} + \dots + G_n G_0^{(n)}; \quad (\text{A.4})$$

$$\begin{aligned} \|G_n^{(n+1)}\| &\leq \|G_0\| + \|G_n^{(n)}\| + \dots + \|G_1^{(n)}\| + \|G_0^{(n)}\| \\ &\leq \|G_0\| \left[1 + \binom{2n-1}{n-1} + \binom{2n-3}{n-1} + \dots + \binom{n}{n-1} + 1 \right] \\ &= \|G_0\| \left[1 + \binom{2n-1}{n} \right] \\ &\leq \|G_0\| \left[\binom{2n-1}{n-1} + \binom{2n-1}{n} \right] = \\ &= \|G_0\| \binom{2n}{n}. \end{aligned}$$

We have shown that $\|G_i^{(n+1)}\| \leq \|G_0\|^{n+1-i} \binom{n+i}{n}$ for $i \leq n$, which proves the result. \square

Lemma A5: Let $G(z)$ be an $m \times m$ inner function such that $\|G_0\| < 1$, with $\|\cdot\|$ any induced matrix norm. Let G^k have a Laurent expansion $G^k(z) = \sum_{i=0}^{\infty} G_i^{(k)} z^{-i}$, and Hankel matrix $\Pi_k := \mathcal{H}(G^k)$. Then for all i

$$\lim_{k \rightarrow \infty} \max_j \|(\Pi_k^* \Pi_k)_{ij} - \delta_{ij}\| = 0.$$

Proof: Consider $R_k(i) = \sum_{t=0}^i \|G_t^{(k)}\|$. With Lemma A4 it follows that $R_k(i) \leq$

$$\begin{aligned} \sum_{t=0}^i \binom{k+t-1}{k-1} \|G_0\|^{k-t} &\leq \left[\sum_{t=0}^i \binom{k+t-1}{k-1} \right] \|G_0\|^{k-i} \\ &= \binom{k+i}{k} \|G_0\|^{k-i} \leq (k+1)^i \|G_0\|^{k-i}. \end{aligned}$$

Since $\|G_0\| < 1$, this implies that $R_k(i) \rightarrow 0$ for $k \rightarrow \infty$.

Now consider the (i, j) -block element of $\Pi_k^* \Pi_k$ with $j \geq i$. $(\Pi_k^* \Pi_k)_{ij} = \sum_{s=0}^{\infty} (G_{i+s}^{(k)})^* G_{j+s}^{(k)} = \delta_{ij} I_m - \sum_{s=0}^{i-1} (G_s^{(k)})^* G_{s+j-i}^{(k)}$. Consequently $\|(\Pi_k^* \Pi_k)_{ij} - \delta_{ij} I_m\| = \|\sum_{s=0}^{i-1} (G_s^{(k)})^* G_{s+j-i}^{(k)}\| \leq \sum_{s=0}^{i-1} \|G_s^{(k)}\| = R_k(i-1) \rightarrow 0$ for $k \rightarrow \infty$.

For $j < i$ it holds that $R_k(j-1) < R_k(i-1)$, which implies that for all j , $\|(\Pi_k^* \Pi_k)_{ij} - \delta_{ij}\| \leq R_k(i-1)$. \square

Proof of Proposition 4.1: Denote $L_i = \sum_{k=0}^{\infty} G_{k+i}^T G_k$ for $i \in \mathbb{Z}$, with $G_j := 0$, $j < 0$. Then $G^T(z^{-1}) G(z) = \sum_{i=-\infty}^{\infty} G_i^T z^i \sum_{k=0}^{\infty} G_k z^{-k}$. This expression equals $\sum_{j=-\infty}^{\infty} (\sum_{k=0}^{\infty} G_{k+j}^T G_k) z^j = \sum_{j=-\infty}^{\infty} L_j z^j$. Since G is inner, $G^T(z^{-1}) G(z) = I$, and evaluation of the former expression for $j \geq 0$ proves the result.

Proof of Proposition 4.3: Part i) follows directly from Proposition 4.1. For Part ii), consider $T_v(\mathcal{H}(G))^*$. Applying Proposition 4.1, shows that $T_v(\mathcal{H}(G))^* = 0$, which implies

that $T_v V_0 U_0^* = 0$ and $T_v V_0 U_0^* U_0 = 0$, leading to $T_v V_0 = 0$. The proof for T_u follows analogously, employing the fact that $G^T(z)$ is inner too. \square

Proof of Lemma 4.4: If G is inner, then for any $k > 1$, $(G^k)^T(z^{-1}) G^k(z) = (G^{k-1})^T(z^{-1}) G^T(z^{-1}) G(z) G^{k-1}(z) = (G^{k-1})^T(z^{-1}) G^{k-1}(z)$, and by induction it follows that G^k is inner. A proof for the McMillan degree of G^k is contained in the proof of Proposition 5.4. \square

Proof of Theorem 4.5:

Part A: A constructive proof will be given in three steps.

- i) The choices for U_j and V_j^* as in (17), (18) lead to matrices Γ_k^o and Γ_k^c in (14), (15), that are unitary;
- ii) The constructed matrix $\Gamma_k^o \Gamma_k^c$ has a block Hankel structure;
- iii) $\Gamma_k^o \Gamma_k^c = \mathcal{H}(G^k)$.

Proofs:

- i) Note that the (i, j) -block-element of $\Gamma_k^o (\Gamma_k^c)^*$ equals $V_0^* T_v^{i-1} (T_v^*)^{j-1} V_0$. With Proposition 4.3 it follows that this equals I for $i = j$, and zero elsewhere.
- ii) This proof will be given by complete induction. For $k = 1$ the statement is true by definition. Assume that it holds for $k - 1$, i.e., $\Pi_{k-1} := \Gamma_{k-1}^o \Gamma_{k-1}^c$ is a Hankel matrix. We have to show that Π_k is a Hankel matrix too, with $\Pi_k = [U_{k-1} \quad \Gamma_{k-1}^o] \begin{bmatrix} \Gamma_{k-1}^c \\ V_{k-1}^* \end{bmatrix}$.

The Markov parameters of the system $G^{k-1}(z)$ will be denoted by H_0, H_1, H_2, \dots .

With U_j and V_j^* chosen as in (17), (18), it follows that

$$\Pi_k = [T_u \Gamma_{k-1}^o \quad U_0] \begin{bmatrix} \Gamma_{k-1}^c \\ V_{k-1}^* \end{bmatrix}$$

showing that $\Pi_k = T_u \Pi_{k-1} + U_0 V_0^* T_v^{k-1}$.

The matrix Π_k has a block Hankel structure if and only if $S \Pi_k = \Pi_k S^*$, with

$$S = \begin{bmatrix} 0 & I & 0 & 0 & \dots \\ 0 & 0 & I & 0 & \dots \\ \vdots & \vdots & 0 & \ddots & \vdots \end{bmatrix}.$$

Evaluation of $S \Pi_k$ shows that $S \Pi_k = S T_u \Pi_{k-1} + S U_0 V_0^* T_v^{k-1} =$

$$= \begin{bmatrix} G_1 \\ G_2 \\ \vdots \end{bmatrix} \Pi_{k-1} + \begin{bmatrix} G_2 & G_3 & \dots \\ G_3 & G_4 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} T_v^{k-1} \quad (\text{A.5})$$

$$= \begin{bmatrix} G_1 \\ G_2 \\ \vdots \end{bmatrix} \begin{bmatrix} H_1 & H_2 & H_3 & \dots \\ \Pi_{k-1} S^* & & & \end{bmatrix} + \begin{bmatrix} G_2 & G_3 & \dots \\ G_3 & G_4 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} T_v^{k-1} \\ = T_u \Pi_{k-1} S^* + \begin{bmatrix} G_1 & G_2 & \dots \\ G_2 & G_3 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} H_1 & H_2 & H_3 & \dots \\ & T_v^{k-1} & & \end{bmatrix}.$$

(A.6)

From Lemma A1 we can deduct that the i th block row of T_v^{k-1} corresponds to the Markov parameters of the transfer function $z^{-i+1} G^{k-1}(z)$. So

$$T_v^{k-1} = \begin{bmatrix} H_0 & H_1 & H_2 & \dots \\ 0 & H_0 & H_1 & \dots \\ \vdots & 0 & H_0 & \dots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix}.$$

As a result (A.6) can be written as

$$S \Pi_k = T_u \Pi_{k-1} S^* + U_0 V_0^* T_v^{k-1} S^* = \Pi_k S^*$$

which proves that Π_k is a block Hankel matrix.

- iii) The proof follows by induction, similarly as in step ii). Consider the first block row of $\Pi_k = \Gamma_k^o \Gamma_k^c$. This equals

$$G_0 [H_1 \quad H_2 \quad \dots] + [G_1 \quad G_2 \quad \dots] T_v^{k-1}.$$

Lemma A1 shows that this is equivalent to

$$G_0 [H_1 \quad H_2 \quad \dots] + [W_0 \quad W_1 \quad \dots]$$

where W_i is such that

$$\sum_{k=0}^{\infty} W_k z^{-k} = \left[\sum_{j=0}^{\infty} G_{j+1} z^{-j} \right] G^{k-1}(z) = \left[\sum_{j=0}^{\infty} G_{j+1} z^{-j} \right] \left[\sum_{i=0}^{\infty} H_i z^{-i} \right].$$

Hence the first block row of Π_k corresponds to the Markov parameters of $\left[\sum_{j=0}^{\infty} G_{j+1} z^{-j} \right] \left[\sum_{i=0}^{\infty} H_i z^{-i} \right] = G^k(z)$.

Part B: Since $\Pi_k = [U_{k-1} \quad \Gamma_{k-1}^o] \begin{bmatrix} \Gamma_{k-1}^c \\ V_{k-1}^* \end{bmatrix}$ is an svd of Π_k , it follows that $U_0^* \Pi_k = V_{k-1}^*$, and $\Pi_k V_0 = U_{k-1}$ which shows the uniqueness of U_{k-1} and V_{k-1} for a given Γ_{k-1}^o and Γ_{k-1}^c . Since U_0, V_0 are unique up to unitary postmultiplication, this holds for the whole sequence of matrices $\{U_i, V_i\}_{i=0,1,\dots}$.

Part C: The proof is given by construction in part a). \square

Proof of Proposition 4.6: With $V_k^* = V_{k-1}^* T_v$ the result follows immediately from Lemma A1. \square

Proof of Theorem 4.7: The result of the theorem follows if the set of basis functions is complete in ℓ_2 , i.e., if for any $x \in \ell_2[0, \infty)$, the following implication holds

$$(\langle \phi_k, x \rangle = 0 \text{ for all } k) \Rightarrow x = 0$$

with $\langle \cdot, \cdot \rangle$ the inner product in ℓ_2 .

If $\langle \phi_k, x \rangle = 0$ for all k , then $(\Gamma_k^c)^* \Gamma_k^c y = 0$ for all k , with $y := [x(0) \ x(1) \ \dots]^*$. Consider the i th row of this equation: $[(\Gamma_k^c)^* \Gamma_k^c]_{i*} y = 0$ for all k , with $[(\Gamma_k^c)^* \Gamma_k^c]_{i*}$ the i th row of the corresponding matrix, then

$$\|[(\Gamma_k^c)^* \Gamma_k^c]_{i*} - e_i^*\| y \leq \|[(\Gamma_k^c)^* \Gamma_k^c]_{i*} - e_i^*\| \cdot \|y\|.$$

Since, $(\Gamma_k^c)^* \Gamma_k^c = (\mathcal{H}(G^k))^* \mathcal{H}(G^k)$, it follows from Lemma A5 that $\lim_{k \rightarrow \infty} \|[(\Gamma_k^c)^* \Gamma_k^c]_{i*} - e_i^*\| = 0$, which implies that $y_i = 0$. \square

Proof of Corollary 4.10: Part a) follows directly from the completeness of the basis. For part b), consider the i th row of $H(z) - D$, with $D = \lim_{z \rightarrow \infty} H(z)$, and $H(z) - D$ written as $\sum_{k=1}^{\infty} h^T(k) z^{-k}$, with $h(k) \in \mathbb{R}^m$.

Consider the scalar time series $\{w(t)\}_{t=0,1,\dots}$ defined by

$$[w(0) \ w(1) \ w(2) \ \dots] = [h^T(1) \ h^T(2) \ \dots].$$

Applying part a) delivers $w(t) = \sum_{k=0}^{\infty} W_k^T \phi_k(t)$, with $W_k \in \mathbb{R}^n$, $\phi_k(t) \in \mathbb{R}^n$. As a result $h^T(j+1) = \sum_{k=0}^{\infty} W_k^T [\phi_k(mj+1) \ \dots \ \phi_k(m(j+1))]$.

In the notation of Proposition 4.6 this leads to $h^T(j+1) = \sum_{k=0}^{\infty} W_k^T M_k(j)$. Consequently $\sum_{i=1}^{\infty} h^T(i) z^{-i} = \sum_{i=1}^{\infty} \sum_{k=0}^{\infty} W_k^T M_k(i-1) z^{-i} = z^{-1} \sum_{k=0}^{\infty} W_k^T V_k(z)$. Since this applies to each row of $H(z) - D$, this proves the result.

Proof of Lemma 5.1: From Proposition 4.2 it follows that for the realization of an inner function, the controllability and observability grammians have to satisfy $PQ = I$, while stability requires that $P, Q \geq 0$. In a balanced realization $P = Q$ and diagonal, which implies $P = Q = I$. \square

Proof of Proposition 5.2:

$$\begin{aligned} G^T(z^{-1})G(z) &= [B^*(z^{-1}I - A^*)^{-1}C^* + D^*] \\ &\quad \cdot [C(zI - A)^{-1}B + D] \\ &= B^*(z^{-1}I - A^*)^{-1}C^*C(zI - A)^{-1}B + \\ &\quad + D^*C(zI - A)^{-1}B \\ &\quad + B^*(z^{-1}I - A^*)^{-1}C^*D + D^*D. \quad (\text{A.7}) \end{aligned}$$

Using $A^*A + C^*C = I$, we can rewrite the first term of the right-hand side by employing $I - A^*A = A^*(zI - A) + (z^{-1}I - A^*)A + (z^{-1}I - A^*)(zI - A)$.

Substitution of this in (A.7) shows that

$$\begin{aligned} G^T(z^{-1})G(z) &= (D^*C + B^*A)(zI - A)^{-1}B \\ &\quad + B^*(z^{-1}I - A^*)^{-1}(C^*D + A^*B) \\ &\quad + B^*B + D^*D. \end{aligned}$$

Since (A, B) is a controllable pair, it follows that $G^T(z^{-1})G(z) = I$ if and only if $B^*B + D^*D = I$ and $D^*C + B^*A = 0$. \square

Proof of Proposition 5.3: Using Proposition 5.2, and its dual version, it follows that $DD^* + CC^* = D^*D + B^*B = I$. Now $\|D\|_2 < 1$ is equivalent to the smallest singular value of B^*B being greater than zero which is equivalent to $\text{rank } B = m$. The result for $\text{rank } C$ follows analogously. \square

Proof of Proposition 5.4: We use complete induction on k to prove this proposition. Note that we can write

$$\begin{aligned} A_k &= \begin{bmatrix} A_{k-1} & 0 \\ BC_{k-1} & A \end{bmatrix} & B_k &= \begin{bmatrix} B_{k-1} \\ BD^{k-1} \end{bmatrix} \\ C_k &= [DC_{k-1} \ C] \end{aligned}$$

with $(A_1, B_1, C_1, D_1) = (A, B, C, D)$. Validity of the statement for (A_1, B_1, C_1, D_1) is straightforward. Assuming validity for $k-1$, we have to show that the statement holds for k . First we show that (A_k, B_k, C_k, D_k) is indeed a realization of $G^k(z)$

$$\begin{aligned} C_k(zI - A_k)^{-1}B_k + D_k &= [DC_{k-1} \ C] \cdot \\ &\quad \begin{bmatrix} (zI - A_{k-1})^{-1} & 0 \\ (zI - A)^{-1}BC_{k-1}(zI - A_{k-1})^{-1} & (zI - A)^{-1} \end{bmatrix} \\ &\quad \cdot \begin{bmatrix} B_{k-1} \\ BD^{k-1} \end{bmatrix} + D^k \\ &= [DC_{k-1} + C(zI - A)^{-1}BC_{k-1}](zI - A_{k-1})^{-1}B_{k-1} + \\ &\quad + C(zI - A)^{-1}BD^{k-1} + D^k \\ &= [D + C(zI - A)^{-1}B]C_{k-1}(zI - A_{k-1})^{-1}B_{k-1} + \\ &\quad + [C(zI - A)^{-1}B + D]D^{k-1} \\ &= [D + C(zI - A)^{-1}B][C_{k-1}(zI - A_{k-1})^{-1}B_{k-1} + D_{k-1}] \\ &= G(z)G^{k-1}(z) = G^k(z). \end{aligned}$$

Balancedness of the realization (A_k, B_k, C_k) can be shown by evaluating: $A_k A_k^* + B_k B_k^*$. For brevity of notation, we will write $(A_{k-1}, B_{k-1}, C_{k-1}, D_{k-1}) = (A, B, C, D)$

$$\begin{aligned} A_k A_k^* + B_k B_k^* &= \\ &\quad \begin{bmatrix} AA^* + BB^* & AC^*B^* + BD^*B^* \\ BCA^* + BDB^* & BCC^*B + AA^* + BDD^*B^* \end{bmatrix} \\ &= \begin{bmatrix} AA^* + BB^* & (AC^* + BD^*)B^* \\ B(CA^* + DB^*) & B(CC^* + DD^*)B^* + AA^* \end{bmatrix}. \end{aligned}$$

Employing Proposition 5.2 together with $AA^* + BB^* = I$ shows that the above expression equals the identity matrix. In a similar way, using the dual forms, it can be shown that $A_k^* A_k + C_k^* C_k = I$, which proves that the realization is balanced and minimal. \square

Proof of Proposition 5.5: Since $V_0^* = [B \ AB \ A^2B \ \dots]$ and $\mathcal{H}(G) = U_0 V_0^*$ is an svd, it follows that $U_0^* = [C^* \ A^*C^* \ \dots]$. Similarly it holds for any k that $\mathcal{H}(G^k) = \Gamma_k^o \Gamma_k^c$ is an svd, with $(\Gamma_k^o)^* = [C_k^* \ A_k^*C_k^* \ \dots]$. Since Γ_k^o and Γ_k^c satisfy the recursion property of Theorem 4.5-a), the given solution has to be the unique one. \square

Proof of Proposition 5.6: With $X = BC$ and P any matrix satisfying $PB = BD$, the matrices A_k, B_k as in Proposition 5.4 will take the form

$$\begin{aligned} A_k &= \begin{bmatrix} A & 0 & \dots & 0 \\ X & A & 0 & 0 \\ PX & X & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ P^{k-2}X & P^{k-1}X & \dots & X & A \end{bmatrix} \text{ and} \\ B_k &= \begin{bmatrix} B \\ PB \\ P^2B \\ \vdots \\ P^{k-1}B \end{bmatrix}. \end{aligned}$$

we can write

$$A_k^j B_k = \begin{bmatrix} M_0(j) \\ M_1(j) \\ \vdots \\ M_{k-1}(j) \end{bmatrix} = A_k(A_k^{j-1} B_k) = \\ = A_k \begin{bmatrix} M_0(j-1) \\ M_1(j-1) \\ \vdots \\ M_{k-1}(j-1) \end{bmatrix}.$$

With the above representation of A_k this leads to the recursive relation (30). Relations (29), (31) follow directly from B_k . \square

Proof of Theorem 5.7: 1-a). Using Proposition 4.6 we have to show that

$$z(zI - A)^{-1} B[D + C(zI - A)^{-1} B]^k = \\ = [(zI - A)^{-1} F(I - zA^*)]^k z(zI - A)^{-1} B.$$

Note that it is sufficient to show that this holds for $k = 1$, since successive application of the equality for $k = 1$ shows the result for any k . The equality for $k = 1$ is equivalent to

$$\{B[D + C(zI - A)^{-1} B] = F(I - zA^*)(zI - A)^{-1} B\} \\ \Leftrightarrow \{BD + BC(zI - A)^{-1} B = BC(zI - A)^{-1} B + \\ - PA(zI - A)^{-1} B - zFA^*(zI - A)^{-1} B\}, \\ \Leftrightarrow \{BD = -PA(zI - A)^{-1} B - zFA^*(zI - A)^{-1} B\}.$$

With $PB = BD$ it suffices to show that

$$P = -PA(zI - A)^{-1} - zFA^*(zI - A)^{-1}.$$

This is equivalent to $\{P(zI - A) = -PA - zFA^*\} \Leftrightarrow \{P = -FA^*\} \Leftrightarrow \{P = -BCA^* + PAA^*\} \Leftrightarrow \{P = BDB^* + P - PBB^*\} \Leftrightarrow \{PBB^* = BDB^*\}$ which is known to be true since $PB = BD$.

1-b) This follows directly from Proposition 4.6.

2) Take $P = -RA^*$. Then $PB = -RA^*B$, which with Proposition 5.2-i) equals RC^*D . With $B = RC^*$ it follows that $PB = BD$ and thus this choice of P satisfies (28).

Now it has to be shown that for this P , $BC - PA = R$. This follows from $BC - PA = BC + RA^*A = BC + R(I - C^*C) = BC + R - BC = R$. \square

Proof of Theorem 5.8: Denote the infinite-dimensional matrices $B_\infty := B_k$, $k \rightarrow \infty$, and $A_\infty := A_k$, $k \rightarrow \infty$. Then we can rewrite (34) as

$$[C_s B_s \ C_s A_s B_s \ C_s A_s^2 B_s \ \dots] = \\ = [L_0 \ L_1 \ \dots][B_\infty \ A_\infty B_\infty \ A_\infty^2 B_\infty \ \dots]. \quad (A.8)$$

Because of the orthonormality of $[B_\infty \ A_\infty B_\infty \ \dots]$ postmultiplication of (A.8) with $[B_\infty \ A_\infty B_\infty \ \dots]^*$ provides

$$[C_s B_s \ C_s A_s B_s \ \dots][B_\infty \ A_\infty B_\infty \ \dots]^* = [L_0 \ L_1 \ L_2 \ \dots]$$

leading to $[L_0 \ L_1 \ L_2 \ \dots] = \sum_{k=0}^{\infty} C_s A_s^k B_s B_\infty^* (A_\infty^*)^k$.

We define

$$Q := [Q_0 \ Q_1 \ Q_2 \ \dots] = \sum_{k=0}^{\infty} A_s^k B_s B_\infty^* (A_\infty^*)^k \quad (A.9)$$

and as a result, $L_k = C_s Q_k$, which equals (35).

Based on (A.9) we can write $A_s Q A_\infty^* = Q - B_s B_\infty^*$. With $X = BC$ and P any matrix satisfying $PB = BD$, this leads to

$$A_s [Q_0 \ Q_1 \ Q_2 \ \dots] \begin{bmatrix} A^* & X^* & X^* P^* & \dots & \dots \\ 0 & A^* & X^* & X^* P^* & \dots \\ 0 & 0 & A^* & X^* & \dots \\ \vdots & \vdots & 0 & A^* & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \\ = [Q_0 \ Q_1 \ Q_2 \ \dots] - B_s [B^* \ B^* P^* \ B^* (P^*)^2 \ \dots].$$

The first element of this equation shows $A_s Q_0 A^* = Q_0 - B_s B^*$, which equals (36). The i th element leads to

$$A_s [Q_i A^* + \sum_{j=1}^i Q_{i-j} X^* (P^*)^{j-1}] = Q_i - B_s B^* (P^*)^i. \quad (A.10)$$

Postmultiplication with P^* gives

$$A_s [Q_i A^* P^* + \sum_{j=1}^i Q_{i-j} X^* (P^*)^j] = Q_i P^* - B_s B^* (P^*)^{i+1}. \quad (A.11)$$

Writing (A.10) for $i \rightarrow i+1$, shows that

$$A_s [Q_{i+1} A^* + \sum_{j=1}^{i+1} Q_{i-j+1} X^* (P^*)^{j-1}] = Q_{i+1} - B_s B^* (P^*)^{i+1} \quad (A.12)$$

and subtracting (A.11) from (A.12) delivers

$$A_s Q_{i+1} A^* - A_s Q_i [A^* P^* - X^*] = Q_{i+1} - Q_i P^*. \quad (A.13)$$

Note that $P = P(AA^* + BB^*) = PAA^* + BDB^* = PAA^* - BCA^* = -FA^*$, and since $F = X - PA$, (A.13) leads to (37). \square

Proof of Proposition 7.1: Part a) A similar result for continuous-time systems is proven in [11]. The discrete-time version follows by applying a bilinear transformation, as is shown in [19].

Part b) The proof is based on an equivalence relation as employed in [11], based on the bilinear transformation. If (A_d, B_d, C_d, D_d) is a balanced realization of a square discrete-time system G_d , then (A_c, B_c, C_c, D_c) is a (continuous-time) balanced realization of a continuous-time system G_c , where

$$A_c = [A_d - I][A_d + I]^{-1} \quad A_d = [I + A_c][I - A_c]^{-1} \\ B_c = \sqrt{2}[A_d + I]^{-1} B_d \quad B_d = \sqrt{2}[I - A_c]^{-1} B_c \\ C_c = \sqrt{2} C_d [A_d + I]^{-1} \quad C_d = \sqrt{2} C_c [I - A_c]^{-1} \\ D_c = D_d - C_d [A_d + I]^{-1} B_d \quad D_d = D_c + C_c [I - A_c]^{-1} B_c.$$

Furthermore G_d is inner if and only if G_c is inner. From [11] it follows that G_c is inner if and only if the following conditions are satisfied:

- 1) $A_c + A_c^* + B_c B_c^* = A_c + A_c^* + C_c^* C_c = 0$
- 2) $D_c D_c^* = D_c^* D_c = I$
- 3) $D_c^* C_c + B_c^* = D_c B_c^* + C = 0$.

Given A_d, B_d we can construct A_c, B_c with the equations above, additionally choosing $C_c := B_c^*$, $D_c := -I$ it follows that (A_c, B_c, C_c, D_c) is a balanced realization of an inner function.

Now transforming the continuous-time realization back to the discrete-time domain, with the expressions for C_d, D_d

as given before and employing the relation $(I - A_c)^{-1} = \frac{1}{2}(I + A_d)$, shows that

$$C_d = B_d^*(A_d^* + I)^{-1}(A_d + I) \quad (A.14)$$

$$D_d = B_d^*(A_d^* + I)^{-1}B_d - I \quad (A.15)$$

which completes the balanced realization (A_d, B_d, C_d, D_d) .

Using the fact that B_d and C_d have full rank, Proposition 5.2 now implies that premultiplication of C_d and D_d with any unitary matrix U characterizes the intended class of balanced realizations.

Part c) Consider a realization (A, B, C, D) satisfying (44), (45) with $U = I$. For $\bar{F} = [I + A][I + A^*]^{-1}$ it follows immediate that $\bar{F}C^* = B$, and the result follows with Theorem 5.7-2).

Denote $F(\lambda, U) = [\lambda I + B(U - \lambda I)(B^*B)^{-1}B^*]\bar{F}$, with U a unitary matrix and $\lambda \in \mathcal{C}$. Then by substitution it can be verified that $F(\lambda, U)C^*U^* = B$. This means that for any $\tilde{C} = UC$ we have constructed a λ -family $F(\lambda, U)$ that satisfies $F(\lambda, U)\tilde{C}^* = B$. Again with Theorem 5.7-2) and choosing $\lambda = 1$ this proves the result. \square

Proof of Proposition 8.1: This result follows directly from rewriting the Lyapunov equations in Theorem 5.8 in terms of Kronecker products: see [3], [19]. \square

Proof of Theorem 8.3: Consider a single scalar entry of the rational matrix function $W(z) = Z_o(zI - X_o)^{-1}Y_o$, written as $w(z) = \sum_{k=1}^{\infty} w_k z^{-k}$. Then $w(z)$ is convergent for $|z| > \lambda$. Consequently, according to basic theory of power series, see e.g., [20], as employed also in [17], there exists an $\alpha \in \mathbb{R}$ such that for each $\eta > \lambda$, $|w_k \eta^{-k}| \leq \alpha$, leading to $|w_k| \leq \alpha \eta^k$. Since this holds for any entry of $W(z)$, and $W(z) = \sum_{k=1}^{\infty} \text{Vec}(L_{k-1})z^{-k}$, there exist scalars α_{ij} such that $|L_k(i, j)| \leq \alpha_{ij} \eta^{k+1}$, with $L_k(i, j)$ being the (i, j) -entry in L_k . Denoting $E(z) := H(z) - \hat{H}^N(z) = \sum_{k=N}^{\infty} L_k V_k(z)$, it follows that $\|E(z)\|_{\infty} \leq \|\sum_{k=N}^{\infty} L_k V_k(z)\|_{\infty} \leq \|V_0(z)\|_{\infty} \sum_{k=N}^{\infty} \|L_k\|_{\infty}$.

Using the above upper bound for $|L_k(i, j)|$ together with the well-known relation between the ∞ -norm and the Frobenius norm, $\|\cdot\|_{\infty} \leq \|\cdot\|_F$, it follows that

$$\|L_k\|_{\infty} \leq \eta^{k+1} \sqrt{\sum_{i,j} \alpha_{ij}^2}. \quad (A.16)$$

Substituting this latter upper bound in the derived upper bound for $\|E(z)\|_{\infty}$, the result of the theorem follows with $c = \|V_0(z)\|_{\infty} \sqrt{\sum_{i,j} \alpha_{ij}^2}$. \square

REFERENCES

- [1] N. I. Akhiezer and I. M. Glazman, *Theory of Linear Operators in Hilbert Space*, vol. 1. Boston, MA: Pitman Adv. Publ. Program, 1981.
- [2] L. Baratchart and M. Oliivi, "Inner-unstable factorization of stable rational transfer functions," in *Modeling, Estimation and Control of Systems with Uncertainty*, G.B. DiMasi, A. Gombani and A.B. Kurzhanski, Eds. Boston, MA: Birkhäuser Verlag, 1991, pp. 22-39.
- [3] R. Bellman, *Introduction to Matrix Computations*. New York: McGraw-Hill, 1970.
- [4] P. M. M. Bongers and P. S. C. Heuberger, "Discrete normalized coprime factorization," in *Proc. 9th Int. Conf. Analysis and Optimization of Systems*, Antibes, France, June 12-15, 1990, pp. 307-313.
- [5] C.-C. Chu, "On discrete inner-outer and spectral factorizations," in *Proc. Amer. Contr. Conf.*, Atlanta, GA, 1988, pp. 1699-1700.
- [6] P. R. Clement, "Laguerre functions in signal analysis and parameter identification," *J. Franklin Inst.*, vol. 313, no. 2, pp. 85-95, 1982.
- [7] G. J. Clowes, "Choice of the time scaling factor for linear system approximations using orthonormal Laguerre functions," *IEEE Trans. Automat. Contr.*, vol. AC-10, no. 5, pp. 487-489, 1965.
- [8] C. A. Desoer, R.-W. Liu, J. Murray and R. Seaks, "Feedback system design: the fractional representation approach to analysis and synthesis," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 399-412, 1980.
- [9] G. A. Dumont, Y. Fu, and A.-L. Elshafei, "Orthonormal functions in identification and adaptive control," in *Intelligent Tuning and Adaptive Control, Selected Papers from the IFAC Symposium* Oxford, England: Pergamon Press, 1991, pp. 193-198.
- [10] B. A. Francis, *A Course in H_{∞} Control Theory* (Lecture Notes in Control Information Sciences) vol. 88. Berlin: Springer-Verlag, 1987.
- [11] K. Glover, "All optimal Hankel-norm approximations of linear multivariable systems and their L_{∞} -error bounds," *Int. J. Contr.*, vol. 39, 1115-1193, 1984.
- [12] K. Glover, J. Lam and J. R. Partington, "Rational approximation of a class of infinite-dimensional systems. I. Singular values of Hankel operators," *Math. Contr. Signals Syst.*, vol. 3, no. 4, pp. 325-344, 1990.
- [13] —, "Rational approximation of a class of infinite-dimensional systems. II. Optimal convergence rates of L_{∞} approximants," *Math. Contr. Signals Syst.*, vol. 4, no. 3, pp. 233-246, 1991.
- [14] M. J. Gottlieb, "Concerning some polynomials orthogonal on finite or enumerable set of points," *Amer. J. Math.*, vol. 60, pp. 453-458, 1938.
- [15] G. Gu, P. P. Khargonekar, and E. B. Lee, "Approximation of infinite-dimensional systems," *IEEE Trans. Automat. Contr.*, vol. 34, no. 6, pp. 610-618, 1989.
- [16] S. Gunnarsson and B. Wahlberg, "Some asymptotic results in recursive identification using Laguerre models," *Int. J. Adaptive Contr. Signal Processing*, vol. 5, no. 5, pp. 313-333, 1991.
- [17] A. J. Helmicki, C. A. Jacobson, and C. N. Nett, "Control oriented system identification: a worst-case/deterministic approach in H_{∞} ," *IEEE Trans. Automat. Contr.*, vol. 36, no. 10, pp. 1163-1176, 1991.
- [18] P. S. C. Heuberger and O. H. Bosgra, "Approximate system identification using system based orthonormal functions," in *Proc. 29th IEEE Conf. Decis. Contr.*, Honolulu, HI, 1990, pp. 1086-1092.
- [19] P. S. C. Heuberger, "On approximate system identification with system based orthonormal functions," Ph.D. dissertation, Delft University of Technology, The Netherlands, 1991.
- [20] E. Hille, *Analytic Function Theory*, vol. 1. Boston, MA: Ginn, 1959.
- [21] W. H. Kautz, "Transient synthesis in the time domain," *IRE Trans. Circuit Theory*, vol. CT-1, pp. 29-39, 1954.
- [22] R. E. King and P. N. Paraskevopoulos, "Digital Laguerre filters," *Int. J. Circuit Theory and Appl.*, vol. 5, pp. 81-91, 1977.
- [23] —, "Parametric identification of discrete time SISO systems," *Int. J. Contr.*, vol. 30, pp. 1023-1029, 1979.
- [24] Y. W. Lee, "Synthesis of electrical networks by means of the Fourier transforms of Laguerre functions," *J. Math. Physics*, vol. 11, pp. 83-113, 1933.
- [25] —, *Statistical Theory of Communication*. New York: Wiley, 1960.
- [26] L. Ljung, *System Identification—Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
- [27] P. M. Mäkilä, "Approximation of stable systems by Laguerre filters," *Automatica*, vol. 26, pp. 333-345, 1990.
- [28] —, "Laguerre series approximation of infinite dimensional systems," *Automatica*, vol. 26, no. 6, pp. 985-995, 1990.
- [29] —, "Laguerre methods and H_{∞} identification of continuous-time systems," *Int. J. Contr.*, vol. 53, no. 3, pp. 689-707, 1991.
- [30] M. Masnadi-Shirazi and N. Ahmed, "Optimum Laguerre networks for a class of discrete-time systems," *IEEE Trans. Signal Processing*, vol. 39, no. 9, pp. 2104-2108, 1991.
- [31] Y. Nurges, "Laguerre models in problems of approximation and identification of discrete systems," *Automat. Remote Contr.*, vol. 48, 346-351, 1987.
- [32] Y. Nurges and Y. Yaaksoo, "Laguerre state equations for a multivariable discrete system," *Automat. Remote Contr.*, vol. 42, 1601-1603, 1981.
- [33] R. Ober and D. McFarlane, "Balanced canonical forms for minimal systems: A normalized coprime factor approach," *Linear Algebra Appl.*, vol. 122/123/124, pp. 23-64, 1989.
- [34] P. N. Paraskevopoulos, "System analysis and synthesis via orthonormal polynomial series and Fourier series," *Math. Comput. Simulation*, vol. 27, 453-469, 1985.
- [35] M. Schetzen, "Power-series equivalence of some functional series with applications," *IEEE Trans. Circuit Theory*, vol. CT-17, no. 3, pp. 305-313, 1970.
- [36] —, "Asymptotic optimum Laguerre series," *IEEE Trans. Circuit Theory*, vol. CT-18, no. 5, pp. 493-500, 1971.

- [37] G. Szegő, *Orthogonal Polynomials*, 4th ed. Providence, RI: American Mathematical Soc. 1975.
- [38] H. Unbehauen and G. P. Rao, "Continuous-time approaches to system identification," *Prepr. 8th IFAC/IFORS Symposium Identification and System Param. Estim.*, Beijing, China, 1988, pp. 60-68.
- [39] P. M. J. Van den Hof, P. S. C. Heuberger, and J. Bokor, "Identification with generalized orthonormal basis functions—Statistical analysis and error bounds," *Preprints 10th IFAC Symposium on System Identification*, Copenhagen, vol. 3, 1994, pp. 207-212, to appear in *Automatica*, vol. 31, no. 12.
- [40] M. Vidyasagar, *Control Systems Synthesis: A Factorization Approach*. Cambridge, MA: MIT Press, 1985.
- [41] B. Wahlberg, "On the use of orthogonalized exponentials in system identification," Dept. Electr. Eng., Linköping University, Sweden, Rep. LiTH-ISY-1099, 1990.
- [42] B. Wahlberg and E. J. Hannan, "Parametric signal modeling using Laguerre filters," *Annals Appl. Prob.*, vol. 3, pp. 467-496, 1993.
- [43] B. Wahlberg, "System identification using Laguerre models," *IEEE Trans. Automat. Contr.*, vol. 36, 551-562, 1991.
- [44] ———, "System identification using Kautz models," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 1276-1282, 1994.
- [45] E. E. Ward, "The calculation of transients in dynamical systems," *Proc. Cambridge Philos. Soc.*, vol. 50, pp. 49-59, 1954.
- [46] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. Cambridge, MA: MIT Press, 1949.
- [47] C. Zervos, P. R. Bélanger and G.A. Dumont, "On PID controller tuning using orthonormal series identification," *Automatica*, vol. 24, no. 2, pp. 165-175, 1988.



Peter S. C. Heuberger was born in Maastricht, The Netherlands, in 1957. He obtained the M.Sc. degree in Mathematics from the Groningen State University in 1983 and the Ph.D. degree in 1991 from Delft University of Technology.

From 1983 until 1990 he was a Research Assistant in the Mechanical Engineering Systems and Control Group at Delft University of Technology. Since 1991, he has been a staff member of the Dutch National Institute of Public Health and Environmental Protection in Bilthoven, The Netherlands, dealing

with modeling and calibration of environmental systems. His main research interests are in issues of system identification and model reduction.



Paul M. J. Van den Hof (S'85-M'88) was born in Maastricht, The Netherlands, in 1957. He obtained the M.Sc. and Ph.D. degrees both from the Department of Electrical Engineering, Eindhoven University of Technology, The Netherlands, in 1982 and 1989, respectively.

From 1986 to 1990 he was an Assistant Professor in the Mechanical Engineering Systems and Control Group at Delft University of Technology, The Netherlands. Since 1991 he has been working in this same group as an Associate Professor. In 1992 he has held a short term visiting position at the Centre for Industrial Control Science, The University of Newcastle, NSW, Australia. His research interests are in issues of system identification, parameterization, and the interplay between identification and robust control design, with applications in mechanical servo systems and industrial process control systems.

Dr. Van den Hof is an Associate Editor of *Automatica*. For his M.Sc. thesis he received the 1983 Control Systems Award (*Regeltechniekprijs*) of the Royal Dutch Institute of Engineers (KIVI).



Okko H. Bosgra was born in Groningen, The Netherlands, in 1944. He obtained the M.Sc. degree in mechanical engineering from Delft University of Technology in 1968.

From 1981-1985 he held a Professorship in the Department of Physics at the Agricultural University of Wageningen, The Netherlands. Since 1986 he is a Full Professor in control engineering, heading the Mechanical Engineering Systems and Control Group of Delft University of Technology, The Netherlands. His current research interests are in the

theory of robust identification and control design and their application to industrial problems in the process control field and in the field of mechanical servo motion control.

Prof. Bosgra is a member of the EUCA (European Union Control Association) Governing Board, a founding member of the Dutch Systems and Control Theory Network, and an Editor-At-Large of the new *European Journal of Control*.

H_∞ Control via Measurement Feedback for General Nonlinear Systems

Alberto Isidori, *Fellow, IEEE*, and Wei Kang, *Member, IEEE*

Abstract—This paper shows how the problem of (local) disturbance attenuation via measurement feedback, with internal stability, can be solved for a nonlinear system of rather general structure. The solution of the problem is shown to be related to the existence of solutions of a pair of Hamilton–Jacobi inequalities in n independent variables, which are associated with state-feedback and, respectively, output-injection design. The results of the paper extend a number of recent achievements in this area.

1. INTRODUCTION

THE development of a systematic analysis of the nonlinear equivalent of the H_∞ (sub)optimal control problem was initiated by the important contributions of Ball–Helton [2], Basar–Bernhard [4], and Van der Schaft [10]. In particular, [10] has shown that the solution of the problem in question in the case of full information configuration, that is when the set of measured variables which are available for feedback includes the state of the controlled plant and the exogenous disturbance input, can be determined from the solution of a Hamilton–Jacobi equation, which is the nonlinear version of the Riccati equation considered in analysis of the H_∞ (sub)optimal control problem for linear systems. More recent contributions to this area of research are the works [11], [6], [3], and [7]. In particular, [6] presents a set of sufficient conditions for the solution of the problem of (local) disturbance attenuation in the case of measurement feedback, that is when the set of measured variables is just a function of the state of the plant and of the disturbance input. The paper [3], among many other important contributions, discusses a number of issues related to the necessity of the (sufficient) conditions proposed in [10] and [6]. The paper [7] considers, as the paper [6] does, systems modeled by equations which are affine in both control and disturbance inputs, and shows that, if a local solution is sought, the sufficient conditions given in [6] can be simplified and turned into conditions involving two Hamilton–Jacobi inequalities in n independent variables, associated with the design of a “state-feedback” and, respectively, of an “output-injection gain.”

In the present paper we study the H_∞ (sub)optimal control problem for systems modeled by equations which are not necessarily affine in the inputs, a more general class, nonlinear systems already considered in [3] (which discusses necessary conditions and a separation principle in the case of measurement feedback) and in [12] (which discusses the case of state feedback). More precisely, we present a (more general) necessary condition for the existence of a solution to the problem in the case of measurement feedback and we show that, if this condition is strengthened in a suitable way, then the construction of a feedback law yielding local disturbance attenuation with internal stability becomes possible. The analysis of the necessity extends the results of [3] while the analysis of the sufficiency extends the results of [7].

Consider a nonlinear system modeled by equations of the form

$$\begin{aligned} \dot{x} &= F(x, w, u) \\ z &= Z(x, u) \\ y &= Y(x, w). \end{aligned} \quad (1)$$

The first equation of this system describes a plant with state x , defined on a neighborhood X of the origin in \mathbb{R}^n with control input $u \in \mathbb{R}^m$ and subject to a set of exogenous input variables $w \in \mathbb{R}^r$ which includes disturbances (to be rejected) and/or references (to be tracked). The second equation defines a penalty variable $z \in \mathbb{R}^s$, which may include a tracking error, as well as a cost of the input u needed to achieve the prescribed control goal. The third equation defines a set of measured variables $y \in \mathbb{R}^p$, which are functions of the state plant x and the exogenous input w . The mappings $F(x, w, u)$, $Z(x, u)$, and $Y(x, w)$ are smooth mappings (i.e., mappings of class C^k for some sufficiently large k) defined in a neighborhood of the origin in $\mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R}^m$. We assume also that $F(0, 0, 0) = 0$, $Z(0, 0) = 0$ and $Y(0, 0) = 0$.

The control action to (1) is to be provided by a controller which processes the measured variable y and generates the appropriate control input u , and is modeled by equations of the form

$$\begin{aligned} \dot{\xi} &= \eta(\xi, y) \\ u &= \theta(\xi, y) \end{aligned} \quad (2)$$

in which ξ is defined on a neighborhood Ξ of the origin in \mathbb{R} and $\eta: \Xi \times \mathbb{R}^p \rightarrow \mathbb{R}^n$, $\theta: \Xi \times \mathbb{R}^p \rightarrow \mathbb{R}^m$ are smooth functions satisfying $\eta(0, 0) = 0$ and $\theta(0, 0) = 0$.

The purpose of the control is twofold: to achieve closed-loop stability and to attenuate the influence of the exogenous input w on the penalty variable z . Following numerous authors (see

Manuscript received February 12, 1993; revised September 24, 1993 and April 13, 1994. Paper recommended by Associate Editor, A. Saberi. This work was supported in part by MURST, by the National Science Foundation under Grant ECS-9208306, and by the Air Force Office of Scientific Research under Grant 91-0266.

A. Isidori is with the Department of Systems Sciences and Mathematics, Washington University, St. Louis, MO 63130 and the Dipartimento di Informatica e Sistemistica, Università di Roma “La Sapienza,” 00184 Rome, Italy.

W. Kang is with the Department of Systems Sciences and Mathematics, Washington University, St. Louis, MO 63130 USA.

IEEE Log Number 9408275.

e.g., [13], [5], [3], [11]) the property of disturbance attenuation in a nonlinear system can be characterized in the following way. A nonlinear system

$$\begin{aligned}\dot{x} &= f(x, u) \\ y &= h(x)\end{aligned}\quad (3)$$

is said to be locally dissipative near $(x, u) = (0, 0)$, with respect to a given supply rate $s(u, y)$, if there exists a smooth function $V(x)$, which is nonnegative and vanishes at $x = 0$, such that

$$\frac{\partial V}{\partial x} f(x, u) + s(u, h(x)) \leq 0 \quad (4)$$

for all (x, u) in a neighborhood of $(0, 0)$ (note that there is some abuse of terminology in this definition: following [13], one should say that a system is "dissipative," with respect to the supply rate $s(x, y)$, if the integral version of the inequality (4) holds, which would not require $V(x)$ to be differentiable).

If (3) is locally asymptotically stable (at the equilibrium $x = 0$) and locally dissipative with respect to the supply rate $s(u, y) = \gamma^2 \|u\|^2 - \|y\|^2$, then its output response to any sufficiently small input, from the initial state $x(0) = 0$, satisfies

$$0 \leq V(x(t)) \leq \int_0^t (\gamma^2 \|u(s)\|^2 - \|y(s)\|^2) ds$$

for all $t > 0$. As a consequence (3) has an L_2 gain which is less than or equal to γ (see [5]).

In what follows, we address the problem of finding a controller which renders the closed-loop system (1)–(2) locally asymptotically stable [at the equilibrium $(x, \xi) = (0, 0)$] and locally dissipative with respect to the supply rate $s(w, z) = \gamma^2 \|w\|^2 - \|z\|^2$. This problem will be referred to as the problem of local disturbance attenuation with internal stability.

II. DISTURBANCE ATTENUATION VIA STATE FEEDBACK

As pointed out by various authors (see, e.g., [2], [4]), the problem of finding a feedback law of the form $u = \alpha(x)$ rendering the closed-loop system

$$\begin{aligned}\dot{x} &= F(x, w, \alpha(x)) \\ z &= Z(x, \alpha(x))\end{aligned}\quad (5)$$

(locally) dissipative with respect to the supply rate $s(w, z) = \gamma^2 \|w\|^2 - \|z\|^2$, can be cast as a two player, zero sum, differential game, in which the minimizing player controls the input u and the maximizing player controls the input w .

The Hamiltonian function associated with the game in question in the present case is a function $H: \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R}^m \rightarrow \mathbb{R}$ defined as

$$H(x, p, w, u) = p^T F(x, w, u) + \|Z(x, u)\|^2 - \gamma^2 \|w\|^2.$$

Suppose plant (1) satisfies the following hypothesis.

Assumption A1: The penalty map $Z(x, u)$ is such that the matrix

$$D_1 = \frac{\partial Z}{\partial u}(0, 0)$$

has rank m .

Then, it is easy to see that, in a neighborhood of the point $(x, p, w, u) = (0, 0, 0, 0)$, the function $H(x, p, w, u)$ has unique local saddle point in (w, u) for each (x, p) . More precisely, there exists unique smooth functions $w_*(x, p)$ and $u_*(x, p)$, defined in a neighborhood of $(0, 0)$, satisfying

$$\begin{aligned}\frac{\partial H}{\partial w}(x, p, w_*(x, p), u_*(x, p)) &= 0, \\ \frac{\partial H}{\partial u}(x, p, w_*(x, p), u_*(x, p)) &= 0, \\ w_*(0, 0) &= 0, \quad u_*(0, 0) = 0\end{aligned}$$

and such that

$$\begin{aligned}H(x, p, w, u_*(x, p)) &\leq H(x, p, w_*(x, p), u_*(x, p)) \\ &\leq H(x, p, w_*(x, p), u)\end{aligned}\quad (6)$$

for each (x, p, w, u) in a neighborhood of $(x, p, w, u) = (0, 0, 0, 0)$. The existence of these functions and (6) can easily be deduced from the observation that $H(x, p, w, u)$, viewed as a function of (w, u) , has a Hessian matrix which, at $(x, p, w, u) = (0, 0, 0, 0)$, is equal to

$$\begin{pmatrix} -2\gamma^2 I & 0 \\ 0 & 2D_1^T D_1 \end{pmatrix}$$

where $D_1^T D_1$ is positive definite by hypothesis.

Let now $V: \mathbb{R}^n \rightarrow \mathbb{R}$ be a smooth function, defined in a neighborhood U of $x = 0$ and such that $V(0) = 0$ and $V_x(0) = 0$, set

$$H_*(x, p) = H(x, p, w_*(x, p), u_*(x, p))$$

$$\alpha_1(x) = w_*(x, V_x^T(x)), \quad \alpha_2(x) = u_*(x, V_x^T(x)) \quad (7)$$

and observe that (6) implies in particular

$$H(x, V_x^T(x), w, \alpha_2(x)) \leq H_*(x, V_x^T(x)).$$

Now, suppose the function $V(x)$ is nonnegative (in which case $V(0) = 0$ implies $V_x(0) = 0$) and renders the inequality

$$H_*(x, V_x^T(x)) \leq 0 \quad (8)$$

satisfied for each x in a neighborhood of zero. Then, set $\alpha(x) = \alpha_2(x)$ in (5). This yields a closed-loop system satisfying

$$V_x(x)F(x, w, \alpha_2(x)) + \|Z(x, \alpha_2(x))\|^2 - \gamma^2 \|w\|^2 \leq 0$$

that is a system which, in a neighborhood of $(x, w) = (0, 0)$, has the required dissipativity property.

Inequality (8) is called a Hamilton–Jacobi–Isaacs inequality. If $V(x)$ is positive definite and appropriate additional hypotheses are satisfied, it can be proven that the feedback law $u = \alpha_2(x)$ is also locally asymptotically stabilizing. These additional hypotheses may take different forms. For instance, as shown in [7] for the case of affine nonlinear systems, one may simply assume that the function $V(x)$ satisfies a strict inequality, i.e., that the left-hand side of (8) is negative for each $x \neq 0$. Or, one may assume that the plant (1) has a suitable "detectability" property. In what follows we describe this second choice.

Assumption A2 Any bounded trajectory $x(t)$ of the system

$$\dot{x}(t) = F(x(t), 0, u(t))$$

satisfying

$$Z(x(t), u(t)) = 0$$

for all $t > 0$, is such that $\lim_{t \rightarrow \infty} x(t) = 0$

Remark Note that this assumption is the nonlinear version of the hypothesis that the system with input u and output z (and $w = 0$) has no finite transmission zeros on the imaginary axis \square

In fact, we have the following result (whose proof, pretty similar to the proof of a corresponding result for affine systems given in [7], is omitted)

Proposition II 1 Consider the system (1) and suppose Assumptions A1 and A2 hold. Suppose there exists a smooth positive definite function $V(r)$, locally defined in a neighborhood of $r = 0$ and such that $V(0) = 0$, which satisfies the Hamilton-Jacobi-Isaacs inequality (8). Then, the feedback law $u = \alpha_2(x)$ solves the problem of local disturbance attenuation with internal stability.

Remark Recalling the interpretation of the problem as a two players, zero sum, differential game, the previous arguments show that if there exists a function $V(r)$ which renders $H_*(r, V_r^T(r)) = 0$, the strategy $u = \alpha_2(r)$ is the best strategy for the minimizing player and the strategy $w = \alpha_1(r)$ is the best strategy for the maximizing player. Since in the present setup, w is a disturbance, the latter can be interpreted as the worst possible disturbance affecting the system \square

III. DISTURBANCE ATTENUATION VIA MEASUREMENT FEEDBACK

A. Preliminaries

If the state x of the plant is not available for measurement then the feedback law proposed in the previous section cannot be directly implemented. Motivated by the results obtained in [7] for affine systems, we consider then a controller in which the feedback law $\alpha_2(x)$ is replaced by a law of the form $\alpha_2(\xi)$, where ξ is an "estimate" of x provided by an appropriate auxiliary dynamics. The latter consists of a copy of the dynamics of x corrected by a term proportional to the error which is induced by the estimation on the measured variable y , that is a system described by an equation of the form

$$\dot{\xi} = F(\xi, w, u) + G(\xi)(y - Y(\xi, w)) \quad (9)$$

where the matrix $G(\xi)$, which will be called the output injection gain, is a matrix to be determined.

Since w is not directly available either, we substitute its actual value by the worst possible one, which, as seen in the previous section, at each time t is a function of the value of the state of the plant at this time and has the expression

$w_*(t) = \alpha_1(x(t))$. Thus, in (9) we set $u = \alpha_2(\xi)$ and $w = \alpha_1(\xi)$. This yields a dynamic feedback law of the form

$$\begin{aligned} \dot{\xi} &= F(\xi, \alpha_1(\xi), \alpha_2(\xi)) + G(\xi)(y - Y(\xi, \alpha_1(\xi))) \\ u &= \alpha_2(\xi) \end{aligned} \quad (10)$$

For convenience, we represent the corresponding closed loop system as

$$\dot{x}' = F'(x', w), \quad z = Z'(x') \quad (11)$$

with x' and $F'(x', w)$, as shown at the bottom of the page and $Z'(x') = Z(x, \alpha_2(\xi))$

As before, we try to render this system locally dissipative with respect to the appropriate supply rate, i.e., we seek the existence of a smooth nonnegative function $U(x')$ such that

$$U_{x'} F'(x', w) + \|Z'(x')\|^2 - \gamma^2 \|w\|^2 \leq 0 \quad \text{for all } w \quad (12)$$

In addition, we want the closed-loop system to be locally asymptotically stable. In this respect, it is quite easy to show that if the above dissipation inequality holds and some appropriate conditions are satisfied the closed loop system (11) has an asymptotically stable equilibrium at $x' = 0$. In fact, the following result holds.

Lemma III 1 Consider system (11) and suppose the following

- i) Assumptions A1 and A2 hold
- ii) the system

$$\dot{\xi} = F(\xi, \alpha_1(\xi), 0) - G(\xi)Y(\xi, \alpha_1(\xi)) \quad (13)$$

has a locally asymptotically stable equilibrium at $\xi = 0$,

iii) there exists a smooth function $U(x')$, vanishing at $x' = 0$ and positive for $x' \neq 0$, which satisfies the inequality (12) for $w = 0$.

Then $F'(x', 0)$ has a locally asymptotically stable equilibrium at $x' = 0$.

Proof The positive definite function $U(x')$ satisfies

$$\frac{dU(x'(t))}{dt} \leq -\|Z(x', \alpha_2(\xi))\|^2 < 0$$

and this shows that the closed-loop system is stable at $x' = 0$. To prove asymptotic stability, consider any trajectory $x'(t)$ of $x' = F'(x', 0)$ yielding $U(x'(t)) = 0$ for all $t > 0$. Then the previous inequality implies that

$$Z(x(t), \alpha_2(\xi(t))) = 0 \quad \text{for all } t \geq 0$$

From Assumption A2, we conclude that $\lim_{t \rightarrow \infty} x(t) = 0$. Moreover, Assumption A1 implies that there is a unique smooth function $u = u(x)$, defined in a neighborhood of $x = 0$ such that

$$Z(x, u(x)) = 0 \quad \text{and} \quad u(0) = 0$$

Therefore, $\lim_{t \rightarrow \infty} \tau(t) = 0$ and $Z(x(t), \alpha_2(\xi(t))) = 0$ implies $\lim_{t \rightarrow \infty} \alpha_2(\xi(t)) = 0$

$$x' = \begin{pmatrix} x \\ \xi \end{pmatrix}, \quad F'(x', w) = \begin{pmatrix} F(x, w, \alpha_2(\xi)) \\ F(\xi, \alpha_1(\xi), \alpha_2(\xi)) + G(\xi)Y(x, w) - G(\xi)Y(\xi, \alpha_1(\xi)) \end{pmatrix}$$

This shows that the ω -limit set of the trajectory in question entirely contained into the set

$$\mathcal{M} = \{(r, \xi) \mid r = 0, \alpha_2(\xi) = 0\}$$

any initial condition in this set produces a trajectory in which $r(t) = 0$, while $\xi(t)$, which is necessarily a trajectory of (13), is such that $\lim_{t \rightarrow \infty} \xi(t) = 0$ by hypothesis. Thus, by LaSalle's invariance principle, we conclude that $r^* = 0$ is an asymptotically stable equilibrium of $F^e(r^*, 0)$. \square

In view of this results we see that if the gain matrix $G(\xi)$ is such that

a) the inequality (12) holds for some positive definite function $U^e(r^*)$, and

b) system (13) has a locally asymptotically stable equilibrium at $\xi = 0$, then the feedback law (10) solves the problem of disturbance attenuation (with local asymptotic stability). In the next two subsections, we discuss the problem of how to find a matrix $G(\xi)$ such that these two properties hold

B A Necessary Condition

To (partially) motivate the condition that, in the next subsection, will be used to find an expression of the output injection $G(\xi)$ in the feedback law (10), we first describe a necessary condition for the solution of the problem of disturbance attenuation via measurement feedback.

Consider the Hamiltonian function $K: \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ defined as

$$K(r, p, w, y) + p^T F(r, u, 0) - y^T Y(r, u) + \|Z(r, 0)\|^2 - \gamma^2 \|w\|^2 \quad (14)$$

and suppose the plant (1) satisfies the following hypothesis

Assumption A3 The measurement map $Y(r, w)$ is such that the matrix

$$D_2 = \frac{\partial Y}{\partial w}(0, 0)$$

has rank p

Since

$$\left(\frac{\partial^2 K(r, p, w, y)}{\partial w^2} \right)_{(r, p, w, y) = (0, 0, 0, 0)} = -2\gamma^2 I$$

there exists a smooth function $w(r, p, y)$, defined in a neighborhood of $(0, 0, 0)$ such that

$$\left(\frac{\partial K(r, p, w, y)}{\partial w} \right)_{w=w(r, p, y)} = 0 \quad w(0, 0, 0) = 0$$

Moreover, it is also easy to check that

$$\left(\frac{\partial^2 K(r, p, w(r, p, y), y)}{\partial y^2} \right)_{(r, p, y) = (0, 0, 0)} = \frac{1}{2\gamma^2} D_2 D_2^T$$

thus, there exists a smooth function $y_*(r, p)$, defined in a neighborhood of $(0, 0)$ such that

$$\left(\frac{\partial K(r, p, \hat{w}(r, p, y), y)}{\partial y} \right)_{y=y_*(r, p)} = 0 \quad y_*(0, 0) = 0$$

By construction,

$$K(r, p, w, y) \leq K(r, p, w(r, p, y), y) \quad (15)$$

for all (r, p, w, y) in a neighborhood of $(0, 0, 0, 0)$ and

$$K(r, p, \hat{w}(r, p, y), y) \geq K(r, p, w(r, p, y_*(r, p)), y_*(r, p)) \quad (16)$$

for all (r, p, y) in a neighborhood of $(0, 0, 0)$

Finally, set

$$w_{**}(r, p) = w(r, p, y_*(r, p))$$

The functions $w_{**}(r, p)$ and $y_*(r, p)$ thus defined can be used to express a necessary condition for the existence of solutions to the problem of disturbance attenuation via measurement feedback

Theorem III.1 Consider system (1) and suppose Assumption A3 holds. Suppose the problem of local disturbance attenuation is solved by the feedback law

$$\xi = \eta(\xi, y) \quad u = \theta(\xi)$$

and let $U(r, \xi)$ be a positive definite smooth function satisfying

$$(U_\xi(r, \xi) - U_\xi(r, \xi)) \begin{pmatrix} F(r, w, \theta(\xi)) \\ \eta(\xi, Y(r, u)) \end{pmatrix} + \|Z(r, \theta(\xi))\|^2 - \gamma^2 \|w\|^2 < 0 \quad (17)$$

for all (r, ξ, w) in a neighborhood of $(0, 0, 0)$. Then the positive definite function $W(r) = U(r, 0)$ satisfies

$$K(r, W_r^T(r), w_{**}(r, W_r^{-1}(r)), y_*(r, W_r^{-1}(r))) < 0 \quad (18)$$

for each r in a neighborhood of zero

Proof Set $\xi = 0$ in (17) to obtain

$$W_r(r)I(r, w, 0) + U_\xi(r, 0)\eta(0, Y(r, u)) + \|Z(r, 0)\|^2 - \gamma^2 \|u\|^2 \leq 0$$

Since the function $\eta(0, y)$ vanishes at $y = 0$, there exists a vector of smooth functions $P(r, y)$ such that $U_\xi(r, 0)\eta(0, y) = P^T(r, y)y$ and therefore

$$W_r(r)I(r, w, 0) + P^T(r, Y(r, u))Y(r, w) + \|Z(r, 0)\|^2 - \gamma^2 \|u\|^2 \leq 0$$

Choosing $w = w(r, W_r^{-1}(r), y)$, the latter yields

$$K(r, W_r^T(r), w(r, W_r^{-1}(r), y), P(r, Y(r, w(r, W_r^{-1}(r), y)))) \leq 0$$

Now, let $y(r)$ denote the unique solution of

$$y(r) = P(r, Y(r, w(r, W_r^{-1}(r), y(r)))) \quad y(0) = 0$$

and set $y = y(r)$ into the previous inequality, to obtain

$$K(r, W_r^T(r), w(r, W_r^{-1}(r), y(r)), y(r)) \leq 0$$

At this point, (16) shows that $K(r, W_r^T(r), u(r, W_r^{-1}(r), y_*(r, W_r^{-1}(r)))) \leq 0$ and the result follows. \square

Remark In the case of a system in which $I(r, w, 0)$ and $Y(r, w)$ are affine in w , condition (18) reduces to the necessary condition $IAY^T I(r) \leq 0$ of [3]. We observe, however, that the latter was proven under the hypothesis that $\eta(\xi, y)$ is affine in y , while this hypothesis is not required in the proof of (18). On the other hand, (18) is only locally valid while the necessary condition of [3] is globally valid. \square

C. The Design of the Output Injection Gain

In this subsection we show that, under appropriate hypotheses, the feedback gain can be determined from the solution of an inequality involving the Hamiltonian function $K(x, p, w, y)$ introduced in the previous subsection. The inequality in question is somewhat stronger than the inequality (18), whose necessity was proven in the previous subsection, and has the following form

$$K(x, W_x^T(x), w_{**}(x, W_x^T(x)), y_*(x, W_x^T(x))) - H_*(x, V_x^T) < 0. \quad (19)$$

As a matter of fact, we can prove the following result.

Theorem III.2: Consider the system (1) and suppose:

- i) Assumptions A1, A2, and A3 hold,
- ii) inequality (8) has a smooth solution $V(x)$, defined in a neighborhood of $x = 0$, vanishing at $x = 0$ and positive for $x \neq 0$,
- iii) inequality (19) has a smooth solution $W(x)$, defined in a neighborhood of $x = 0$, vanishing at $x = 0$ and positive for $x \neq 0$,
- iv) $W(x) - V(x) > 0$ for all $x \neq 0$,
- v) the Hessian matrix of

$$K(x, W_x^T(x), w_{**}(x, W_x^T(x)), y_*(x, W_x^T(x))) - H_*(x, V_x^T)$$

is nonsingular at $x = 0$ and the equation

$$(W_x(x) - V_x(x))G(x) = y_*^T(x, W_x^T(x)) \quad (20)$$

has a smooth solution $G(x)$.

Then, the problem of local disturbance attenuation with internal stability is solved by the output feedback

$$\begin{aligned} \dot{\xi} &= F(\xi, \alpha_1(\xi), \alpha_2(\xi)) - G(\xi)Y(\xi, \alpha_1(\xi)) + G(\xi)y \\ u &= \alpha_2(\xi) \end{aligned}$$

with $\alpha_1(x)$, $\alpha_2(x)$, and $G(x)$ chosen as in (7) and (20).

Proof: Set $Q(x) = W(x) - V(x)$ and define

$$S(x, w) = Q_x[F(x, w, 0) - G(x)Y(x, w)] + H(x, V_x^T, w, 0) - H_*(x, V_x^T).$$

Then, it is easy to check that

$$\begin{aligned} S(x, w) &= W_x F(x, w, 0) - y_*^T(x, W_x^T)Y(x, w) \\ &\quad - V_x F(x, w, 0) + H(x, V_x^T, w, 0) \\ &\quad - H_*(x, V_x^T) \\ &= W_x F(x, w, 0) - y_*^T(x, W_x^T)Y(x, w) \\ &\quad + \|Z(x, 0)\|^2 - \gamma^2 \|w\|^2 - H_*(x, V_x^T) \\ &= K(x, W_x^T, w, y_*(x, W_x^T)) - H_*(x, V_x^T) \\ &\leq K(x, W_x^T, w_{**}(x, W_x^T), y_*(x, W_x^T)) \\ &\quad - H_*(x, V_x^T) \\ &= x^T M(x)x \end{aligned}$$

where $M(x)$ is a matrix of smooth functions, which is negative definite at $x = 0$.

Now set $U(x^e) = Q(x - \xi) + V(x)$. It can be proven that the function thus defined is such that the two conditions a) and

b) indicated at the end of Section III-A hold, if $G(x)$ is chosen as in (20). As far as condition a) is concerned, observe that

$$\begin{aligned} U_{x^e} F(x^e, w) &+ \|Z^e(x^e)\|^2 - \gamma^2 \|w\|^2 \\ &= Q_x(x - \xi)[F(x, w, \alpha_2(\xi)) - F(\xi, \alpha_1(\xi), \alpha_2(\xi)) \\ &\quad - G(\xi)Y(x, w) + G(\xi)Y(\xi, \alpha_1(\xi))] \\ &\quad + V_x F(x, w, \alpha_2(\xi)) + \|Z(x, \alpha_2(\xi))\|^2 - \gamma^2 \|w\|^2 \\ &\leq Q_x(x - \xi)[F(x, w, \alpha_2(\xi)) - F(\xi, \alpha_1(\xi), \alpha_2(\xi)) \\ &\quad - G(\xi)Y(x, w) + G(\xi)Y(\xi, \alpha_1(\xi))] \\ &\quad + H(x, V_x^T, w, \alpha_2(\xi)) - H_*(x, V_x^T). \end{aligned}$$

Let $L(x, \xi, w)$ denote the expression on the right-hand side of this inequality. It is easy to see that, in a neighborhood of $(x, \xi) = (0, 0, 0)$, there exists a (unique) function $\tilde{w}(x, \xi)$ such that

$$\left(\frac{\partial L(x, \xi, w)}{\partial w} \right)_{w=\tilde{w}(x, \xi)} = 0, \quad \tilde{w}(0, 0) = 0$$

and

$$L(x, \xi, w) \leq L(x, \xi, \tilde{w}(x, \xi))$$

for all (x, ξ, w) in a neighborhood of $(0, 0, 0)$. Thus, in this neighborhood,

$$U_{x^e} F(x^e, w) + \|Z^e(x^e)\|^2 - \gamma^2 \|w\|^2 \leq L(x, \xi, \tilde{w}(x, \xi)).$$

Moreover, it is also possible to show, by means of somewhat lengthy calculations (omitted for reasons of space), that

$$L(x, \xi, \tilde{w}(x, \xi)) = (x - \xi)^T R(x, \xi)(x - \xi) \quad (21)$$

where $R(x, \xi)$, a matrix of smooth functions, is such that $R(0, 0) = M(0)$. Thus $R(x, \xi)$ is negative definite in a neighborhood of $(0, 0)$ and the result [namely, property a)] follows.

To prove that condition b) holds, it suffices to set $w = \alpha_1(x)$ in the definition of $S(x, w)$, to conclude that

$$0 > S(x, \alpha_1(x)) \geq Q_x(F(x, \alpha_1(x), 0) - G(x)Y(x, \alpha_1(x)))$$

thus showing that $Q(x)$ is a Lyapunov function for (13). \square

Remark: The output injection gain $G(x)$ is implicitly defined by means of (20). Obviously, the existence of a smooth solution $G(x)$ of this equation depends on how $W_x(x) - V_x(x)$ and $y_*(x, W_x^T(x))$, which both vanish at $x = 0$, tend to zero as $x \rightarrow 0$. If, for instance, the function $W(x) - V(x)$ has a Hessian matrix which is nonsingular at $x = 0$, then a smooth solution of (20) always exists. In this case, in fact

$$W_x(x) - V_x(x) = x^T R_1(x)$$

where $R_1(x)$ is a matrix which is nonsingular for each x in a neighborhood of $x = 0$. The right-hand side of (20) can be given a similar expression, say

$$y_*^T(x, W_x^T(x)) = x^T L(x)$$

and the equation in question is indeed solved by $G(x) = R_1^{-1}(x)L(x)$. \square

Remark: It may be useful to stress that the proof of Theorem III.2, as far as property a) is concerned, does not just consist of straightforward "second-order" arguments. It

in this respect, the result cannot be viewed as direct consequence of the corresponding result which is known to hold for linear systems. As a matter of fact, the maximum of $\|F(x^e, w) + \|Z^e(x^e)\|^2 - \gamma^2 \|w\|^2$ in w [that is, the function $L(x, \xi, \tilde{w}(x, \xi))$] is a function of (x, ξ) whose second-order terms vanish at $x = \xi$. In other words, the maximum in w of the left-hand side of (12) has a "second-order" approximation which is only semidefinite, and this does not suffice to conclude that (12) holds. The required conclusion is reached only after a stronger relation like (21) is shown, which clearly is not just a property of second-order terms. \square

IV. AN EXAMPLE

In the previous sections, we have shown how to construct a feedback law which solves the problem of disturbance attenuation, using the solutions $V(x)$ and $W(x)$ of the pair of Hamilton–Jacobi–Isaacs inequalities (8)–(19). In this section we illustrate, with the help of an example, how the problem of determining actual solutions of these inequalities can be addressed. The first thing to observe, in this respect, is that—except for some special classes of systems (like, for instance, the systems in which $F(x, w, u)$, $H(x, u)$ and $Y(x, w)$ are affine in w and u)—the inequalities in question are only implicitly defined [because so are the functions $w_*(x)$, $u_*(x)$, $w_{**}(x)$, $y_*(x)$ which characterize the left-hand sides of (8)–(19)]. Thus, the only feasible practical way to implement the results established so far is to try to estimate the solutions of these inequalities by means of appropriate numerical methods.

To this end, we recall that the problem of determining polynomial approximations (of some prescribed degree) for the solution of a Hamilton–Jacobi equation has been addressed and solved by Albrecht in [1] and Lukes in [9]. The results described in these papers were developed for the specific case of the Hamilton–Jacobi equation arising in nonlinear optimal control, but indeed hold (as, e.g., observed in [11]) also in the more general case of the Hamilton–Jacobi–Isaacs equation of a nonlinear differential game, thus also for the inequalities (8) and (19).

The approach in question can be briefly described as follows. Suppose one is interested in the determination of a positive definite function $V: \mathbb{R}^n \rightarrow \mathbb{R}$, defined in a neighborhood of $x = 0$ and vanishing at $x = 0$, such that

$$H(x, V_x^T(x), w_*(x, V_x^T(x)), u_*(x, V_x^T(x))) = \phi(x) \quad (22)$$

for some negative semidefinite or negative definite analytic function $\phi(x)$. Suppose $V(x)$ is analytic and set

$$\begin{aligned} V(x) &= \sum_{d=1}^{\infty} V^{[d+1]}(x) \\ w_*(x) &= w_*(x, V_x^T(x)) = \sum_{d=1}^{\infty} w_*^{[d]}(x) \\ u_*(x) &= u_*(x, V_x^T(x)) = \sum_{d=1}^{\infty} u_*^{[d]}(x) \\ \phi(x) &= \sum_{d=3}^{\infty} \phi^{[d]}(x) \end{aligned}$$

in which the superscript "[d]" means that a function or the components of a vector are homogeneous polynomials of degree d .

Following [1] and [9], it is possible to show that, for any $d > 1$, $V^{[d+1]}(x)$ and $w_*^{[d]}(x)$, $u_*^{[d]}(x)$ depend only on $V^{[2]}(x)$, $w_*^{[1]}(x)$, $u_*^{[1]}(x)$, \dots , $V^{[d]}(x)$, $w_*^{[d-1]}(x)$, $u_*^{[d-1]}(x)$. The function $V^{[2]}(x)$ is determined by a Riccati equation, involving the parameters which characterize the linear approximation of the plant at the equilibrium $(x, u, w) = (0, 0, 0)$, while $V^{[d+1]}(x)$, for $d > 1$, is determined by a linear equation (called homological equation) which involves the parameters of $V^{[2]}(x)$, $w_*^{[1]}(x)$, $u_*^{[1]}(x)$, \dots , $V^{[d]}(x)$, $w_*^{[d-1]}(x)$, $u_*^{[d-1]}(x)$ (see [9], [11], and [8] for more details).

Thus, it is not difficult to setup up a recursive procedure yielding, after d stages, a polynomial approximation of order $d+1$ for $V(x)$ and a polynomial approximation of order d for $u_*(x)$. Note also that the function $\phi(x)$ on the right-hand side of (22) is quite an arbitrary function (the only requirement being its sign-definiteness) and, thus, the extra degrees of freedom associated with the choice of this function can be exploited to the purpose of simplifying the entire solution process.

In the following (elementary) example we show how the solution of a Hamilton–Jacobi inequality can be achieved via polynomial approximations. This example may also help, in our opinion, to dissipate the impression that the analysis of the problem of "local" disturbance attenuation with internal stability relies solely on "first order" arguments. Consider the system

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} xy \\ x^2 + w + u \end{pmatrix}, \quad z = \begin{pmatrix} y - x^2 \\ u \end{pmatrix}.$$

It is easily seen (by means of simple arguments which are omitted for reasons of space) that for any linear feedback law $u = ax + by$, the equilibrium point $(x, y) = (0, 0)$ of the corresponding closed-loop system (with $w = 0$) is unstable. Thus, in particular, the problem of local disturbance attenuation with internal stability cannot be solved by means of a linear state feedback law. The problem in question, however, can be solved by means of a nonlinear state feedback, as shown hereafter.

The Hamilton–Jacobi inequality (8) assumes in this case the form

$$V_x xy + V_y x^2 + y^2 - 2x^2 y + x^4 + \frac{1}{4} \lambda V_y^2 \leq 0 \quad (23)$$

where $\lambda = (1/\gamma^2) - 1$.

Set

$$V(x, y) = V^{[2]}(x, y) + V^{[3]}(x, y)$$

with

$$\begin{aligned} V^{[2]}(x, y) &= ax^2 + by^2 \\ V^{[3]}(x, y) &= cx^2 y + dx y^2 + ex^3 + fy^3. \end{aligned}$$

In the notation introduced before, we obtain on the left-hand side of the Hamilton–Jacobi inequality (23) a function of (x, y) of the form

$$\phi(x, y) = \Phi^{[2]}(x, y) + \Phi^{[3]}(x, y) +$$

in which

$$\Phi^{[2]}(x, y) = (1 + b^2\lambda)y^2$$

$$\Phi^{[3]}(x, y) = (2a + 2b - 2 + \lambda bc)x^2y + 2\lambda bdx y^2 + 3\lambda bfy^3$$

Since $\phi(x, y)$ is required to be sign-definite, we impose $\Phi^{[3]}(x, y) = 0$, i.e., we impose the following constraints on the parameters which characterize $V(x, y)$

$$a = 1 - b - \frac{\lambda}{2}bc, \quad d = f = 0$$

Moreover, to obtain simpler expressions, we set $c = 0$. As a result, we obtain

$$V(x) = ax^2 + by^2 + cx^2y$$

$$\phi(x, y) = (1 + b^2\lambda)y^2 + \left(\frac{\lambda}{4}c^2 + c + 1\right)x^4 + 2cx^2y^2$$

The function $V(x, y)$ is positive definite, and the function $\phi(x, y)$ is negative definite, in a neighborhood of the origin, if the following constraints hold

$$a > 0, \quad b > 0, \quad 1 + b^2\lambda < 0, \quad \left(\frac{\lambda}{4}c^2 + c + 1\right) < 0$$

For any $\lambda < 0$, i.e., for any $\gamma > 1$, these conditions can indeed be satisfied by suitable (sufficiently large) choices of $b > 0$ and $c > 0$. The positive definite function $V(x, y)$ thus found satisfies the strict version Hamilton–Jacobi–Isaacs inequality (8). Thus, according to the results described before, the corresponding feedback law

$$u = \alpha(x) = -\frac{1}{2}V_y = -by - \frac{c}{2}x^2$$

locally asymptotically stabilizes the equilibrium of the corresponding closed-loop system and renders its L_2 gain less than or equal to γ .

REFERENCES

- [1] E. G. Albekht, "On the optimal stabilization of nonlinear systems," *J. Appl. Math. Mech.* vol. 25, pp. 1254–1266, 1962.
- [2] J. A. Ball and J. W. Helton, " H_∞ optimal control for nonlinear plants. Connection with differential games," in *Proc. 28th Conf. Decision and Control*, Tampa, FL, Dec. 1989, pp. 956–962.
- [3] J. Ball, J. W. Helton, and M. L. Walker, " H_∞ control for nonlinear systems via output feedback," *IEEE Trans. Automat. Contr.* vol. 38, pp. 546–559, 1993.
- [4] T. Başar and P. Bernhard, *H_∞ -Optimal Control and Related Minimax Design Problems*. Birkhäuser, 1990.
- [5] D. Hill and H. Moylan, "The stability of nonlinear dissipative systems," *IEEE Trans. Automat. Contr.* vol. AC-21, pp. 708–711, 1976.
- [6] A. Isidori and A. Astolfi, "Disturbance attenuation and H_∞ control via measurement feedback in nonlinear systems," *IEEE Trans. Automat. Contr.* vol. AC-37, pp. 1283–1293, 1992.
- [7] A. Isidori, " H_∞ control via measurement feedback for affine nonlinear systems," *Int. J. Robust Nonlinear Contr.*, 1993.
- [8] W. Kang, P. K. De, and A. Isidori, "Flight control in a windshear via nonlinear H_∞ methods," in *Proc. 31st Conf. Decision Control*, Tucson, AZ, 1992, pp. 1135–1142.
- [9] D. Lukes, "Optimal regulation of nonlinear systems," *SIAM J. Contr.* vol. 7, pp. 75–100, 1969.
- [10] A. J. Van der Schaft, "A state-space approach to nonlinear H_∞ control," *Syst. Contr. Lett.* vol. 16, pp. 1–8, 1991.
- [11] ———, "L₂-gain analysis of nonlinear systems and nonlinear H_∞ control," *IEEE Trans. Automat. Contr.* vol. 37, pp. 770–784, 1992.
- [12] ———, "Complements to nonlinear H_∞ optimal control by state feedback," *IMA J. Math. Contr. Inf.* vol. 9, pp. 245–254, 1992.
- [13] J. C. Willems, "Dissipative dynamical systems," *Arch. Rational Mechanics Anal.* vol. 45, pp. 321–693, 1972.



Alberto Isidori (M'80–SM'85–F'87) was born in Rapallo, Italy, in 1942. He graduated in electrical engineering from the University of Rome in 1965.

Since 1975, he has been Professor of Automatic Control at the University of Rome. Since 1989, he has also been with Washington University in St. Louis, MO. His research interests are in control theory. He is the author of *Nonlinear Control Systems* (Springer-Verlag, 1985 and 1989), *Topics in Control Theory* (with H. Knobloch and D. Flockner) (Birkhäuser, 1993). He is also editor or coeditor of

various conference proceedings and the author of over 120 articles for the most part on the subject of nonlinear feedback design.

Prof. Isidori is currently a member of the IFAC Council and Vice President of the European Union Control Association. He is presently serving or has served in numerous Editorial Boards of major archival journals which include *Automatica*, *IEEE Transactions on Automatic Control*, *Systems and Control Letters*, and *Mathematics of Control Signals and Systems*. In 1983 and in 1991, he received the Outstanding Paper Award from the Control Systems Society of the IEEE for papers published on the *IEEE Transactions on Automatic Control*. In 1993, he received the Outstanding Paper Award from the International Federation of Automatic Control for a paper published on *Automatica*.



Wei Kang (S'91–M'91) was born in Beijing, P. R. China, in 1960. He received the B.S. and M.S. degrees from the Department of Mathematics of Nankai University, China, in 1982 and 1985, respectively, and the Ph.D. degree in mathematics from the University of California at Davis in 1991.

From 1991 to 1994, he was a Visiting Assistant Professor of Systems Science and Mathematics at Washington University. He joined the faculty of the Department of Mathematics at the Naval Postgraduate School as an Assistant Professor in June

1994. His earlier research publications were related to topics in Lie group Lie algebras and differential geometry. Since 1987, his research interest has been changed to nonlinear control theory and its applications in industry. His published results include feedback invariants, nonlinear controller norms, H_∞ control, feedback linearization, and aircraft control and spacecraft control.

Dr. Kang is a member of AMS and SIAM.

Technical Notes and Correspondence

Perturbation Bounds for Root-Clustering of Linear Systems in a Specified Second Order Subregion

W. Bakker, J. S. Luo, and A. Johnson

Abstract—Sufficient bounds for structured and unstructured uncertainties for root-clustering in a specified second order subregion of the complex plane, for both continuous-time and discrete-time systems, are given using the Generalized Lyapunov Theory. Furthermore, for unstructured uncertainties, a still less conservative result is obtained by shifting the center or focus of the subregion along the real axis to the origin and by applying root-clustering to the “shifted eigenvalue” system matrix, which is obtained by shifting the eigenvalues of the system matrix correspondingly.

I. INTRODUCTION

We study the problem of how to guarantee the location of the eigenvalues of a perturbed system matrix in a specified (symmetric) second-order subregion of the complex plane. The perturbation may be structured or unstructured. The subregions discussed are, in the continuous-time case, subregions of the left-half complex plane (LHP) and in the discrete-time case subregions of the unit disk (UD) centered at the origin. Recently, Abdul-Wahab [1] and [2] discussed second-order subregions for discrete-time and continuous time systems, but the results have turned out to be erroneous. The errors have been pointed out by Yedavalli [3] for the continuous-time case and by Bakker and Luo [4] for the discrete-time case. More recently, Yedavalli [5] obtained results for first and second-order subregions using Generalized Lyapunov Theory presented by Gutman and Jury [6].

Conservatism of the paper of Yedavalli [5] is reduced in several ways. First, the norms used to compute the bound are taken at a later stage compared to the bounds of Yedavalli. Next, an important result is, for unstructured uncertainties, that the bound can be improved for subregions with center or focus at $(\alpha, 0)$. This improved bound is obtained by shifting the center or focus of the subregion along the real axis to the origin and computing the bound for the “shifted eigenvalue” matrix. Generally, a large α yields a large improvement. Finally, from Luo *et al.* [7] we are motivated to use the square root of positive definite matrix $Q (= S^T S)$ in the Lyapunov equations; the result can be improved by choosing the S -matrix appropriately. Using a personal computer, the optimal S -matrix can be found by an optimization program in Matlab (e.g., `fminu.m` [9]).

This paper is organized as follows. In Section II the robust stability analysis for state-space models of continuous- and discrete-time systems using the Generalized Lyapunov Theory [6] is described. In Section III we introduce bounds for both continuous and discrete-time systems with unstructured uncertainties. A motivation for “shifting”

is given, followed by the “shifting theorem.” In Section IV, bounds for structured uncertainties are given. For both cases, the bounds can be improved by optimization programs. In Section V, the results are illustrated with examples, followed by two general remarks. We end the paper with conclusions in Section VI.

II. GENERALIZED LYAPUNOV ROOT-CLUSTERING THEORY

Let the perturbed continuous-time system be represented by

$$\dot{\mathbf{x}}(t) = (A + E)\mathbf{x}(t) \quad (2.1a)$$

and let the perturbed discrete-time system be represented by

$$\mathbf{x}(k+1) = (A + E)\mathbf{x}(k) \quad (2.1b)$$

where the state vector is $\mathbf{x}(t)$ or $\mathbf{x}(k) \in \mathbb{R}^n$, the time-invariant system matrix $A \in \mathbb{R}^{n \times n}$ and the perturbation matrix $E \in \mathbb{R}^{n \times n}$. If the structure of the perturbation is known, the substitution

$$E = \sum_{i=1}^m \epsilon_i E_i \quad (2.2)$$

is made. Here ϵ_i ($i = 1, 2, \dots, m$) are time-invariant uncertain parameters which are assumed to lie in the interval around zero ($\epsilon_i \in [-\epsilon, \epsilon]$ where $\epsilon > 0$) and E_i ($i = 1, 2, \dots, m$) are constant matrices determined by the structure of the parameter uncertainties.

Any symmetric second order subregion in the complex plane ($(s$ or $z) = x + iy$) is described by (x, y) satisfying the inequality [6]

$$\Omega_2 = \{(x, y): \gamma_{00} + \gamma_{10}x + \gamma_{20}x^2 + \gamma_{02}y^2 < 0\} \quad (2.3)$$

where γ_{ij} is a real coefficient (note the difference with the boldfaced state vector \mathbf{x} and the real coordinate x). It is clear that this region is symmetrical with respect to the x axis. This inequality describes for example circles, ellipses, left parabolas, etc. Let us now define, as in [6]

$$\begin{aligned} c_{00} &= \gamma_{00}, & c_{10} &= c_{01} = \frac{1}{2}\gamma_{10}, & c_{11} &= \frac{1}{2}(\gamma_{20} + \gamma_{02}), \\ c_{20} &= c_{02} = \frac{1}{4}(\gamma_{20} - \gamma_{02}) \end{aligned} \quad (2.4)$$

so we can present the Lyapunov criterion for root-clustering in a specified second-order subregion Ω_2 [6]: Let $\gamma_{20} + \gamma_{02} \geq 0$ in (2.3) and let the coefficients c_{pq} be given by (2.4). If and only if for a given positive definite Hermitian matrix Q the following generalized Lyapunov equation (GLE)

$$c_{00}P + c_{10}(A^T P + PA) + c_{11}A^T P A + c_{20}((A^T)^2 P + PA^2) = -Q \quad (2.5)$$

has a positive definite solution P , then all the eigenvalues of the nonperturbed system [$E = 0$ in (2.1)] are located inside the defined subregion (2.3). Equation (2.5) can be solved using the Kronecker product; for a detailed description, see [6].

Manuscript received April 19, 1993; revised January 27, 1994 and May 18, 1994.

W. Bakker is with the ASPC-group, Turnkies Process Control, Henry van Antwerpenstraat 38, 3822 XE Amersfoort, The Netherlands.

J. S. Luo is with Process Dynamics and Control, Kramers Laboratory, Delft University of Technology, Prins Bernhardlaan 6, 2628 BW Delft, The Netherlands.

A. Johnson passed away in February, 1994; he was head of the section Process Dynamics and Control at the Kramers Laboratory.

IEEE Log Number 9407237.

III. PERTURBATION BOUNDS FOR UNSTRUCTURED UNCERTAINTIES

In this section new stability robustness bounds for systems with unstructured uncertainties based on the Generalized Lyapunov Root-Clustering Theory presented in Section II are given. It is shown that "shifting," for a large class of subregions, yields improved bounds. These results are organized in the "shifting theorem."

In the next theorem we give less conservative bounds on the norm of the perturbation matrix E so that the eigenvalues are located in the specified subregion. There is no knowledge assumed about the structure of the perturbation.

Theorem 3.1: Suppose

- 1) The perturbed system is described by (2.1).
- 2) The eigenvalues of the nonperturbed system matrix A are located inside a subregion Ω_2 defined by (2.3).
- 3) S^T and S are full rank (invertible) matrices satisfying $Q = S^T S$.
- 4) P is a symmetric positive definite matrix defined by (2.5).

Then the perturbed system matrix $A + E$ has all the eigenvalues located in the subregion Ω_2 described by (2.3), if

$$\|E\| < \mu = \frac{b}{2a} \left(-1 + \sqrt{1 + \frac{4a}{b^2}} \right) \quad (3.1)$$

with

$$a \equiv |c_{11}| \|P\| \|S^{-1}\|^2 + 2|c_{20}| \|S^{-1}\| \|PS^{-1}\| \quad (3.1a)$$

$$\begin{aligned} b \equiv & 2|c_{10}| \|S^{-1}\| \|PS^{-1}\| + 2|c_{11}| \|S^{-1}\| \|PAS^{-1}\| \\ & + 2|c_{20}| \|S^{-1} A^T\| \|PS^{-1}\| \\ & + 2|c_{20}| \|S^{-1}\| \|A^T P S^{-1}\| \end{aligned} \quad (3.1b)$$

and where $\|A\|$ = the spectral norm of the matrix A (the largest singular value of A).

Proof: See Appendix.

Remark 3.1: For first-order subregions, $c_{11} = c_{20} = 0$ (2.3) and (2.4). In that case the upper bound (3.1) is obtained from (A.2) with $\alpha = 0$ to be $\mu = 1/b$.

Shifting of Subregions

The bound μ depends on the matrices A , P , and S and on the coefficients that describe the subregion. Consider the role of the system matrix in the bound (summarized in Fact 1). Of interest are subregions of the form

$$\Omega_2 = \{(x, y): ((x - \alpha)/v)^f + (y/w)^g - u < 0\} \quad (3.2)$$

where $(\alpha, v, w, u) \in \mathbf{R}$ and $v > 0$, $u > 0$; $f = 1, 2$, and $g = 0, 2$. Examples are a circle, a left-parabola, an ellipse, a vertical strip, etc. The eigenvalues of system matrix A have to be located in this subregion. Next, for convenience, restrict the subregion to be a circular subregion of the LHP, with center $(\alpha, 0)$ and radius 1 satisfying $\alpha < -1$, $f = 2 = g$, $u = 1 = v = w$ in (3.2) (for other subregions similar arguments hold). Then the eigenvalues of the matrix A have to be located within a distance 1 from $(\alpha, 0)$ and therefore satisfy

$$(\alpha - 1) < \operatorname{Re}(\lambda(A)) < (\alpha + 1).$$

Because $\alpha < -1 \Leftrightarrow \alpha + 1 < 0$, taking norms

$$|\operatorname{Re}(\lambda(A))| > |\alpha + 1|.$$

From [20]

$$\begin{aligned} \|A\| &\geq |\lambda(A)| = (\operatorname{Re}(\lambda(A))^2 + \operatorname{Im}(\lambda(A))^2)^{1/2} \\ &\geq |\operatorname{Re}(\lambda(A))| > |\alpha + 1|. \end{aligned} \quad (3.3)$$

Fact 1: The coefficients c_{11} and c_{20} and the solution for P (2.5) (see Remark 3.2), used in the bound (3.1a), are not affected by shifting.

Fact 2: From (3.3), $\|A\| > |\alpha + 1|$. This implies, with Fact 1, that when $S = I$ at least the third term in (3.1b) is multiplied by a factor larger than $|\alpha + 1|$. Thus when α is small ($\alpha \ll -1$), $|\alpha + 1|$ is large and consequently that $\|A\|$ is large. So a large α causes a large b . For subregions as a left-parabola or a vertical strip, with center or focus at $(\alpha, 0)$, similar inequalities hold.

Fact 3: The value of a is unaffected by shifting (follows from Fact 1).

Fact 4: When b becomes large (a unaffected) then the bound becomes small (i.e., very conservative). This is even more obvious when, with $b \gg a$, a Taylor series expansion is used for the square root in (3.1): $\mu \approx 1/b$.

We have shown that the bound is conservative when subregions with center or focus at $(\alpha, 0)$, α sufficiently large, are considered. In the next theorem, it is shown that it is possible to shift the subregions (3.2) from $(\alpha, 0)$ to $(0, 0)$ and to compute the bound for the "shifted eigenvalue" system matrix with Theorem 3.1. With the facts are given above, the improvement of the bound should be clear.

Theorem 3.2 "Shifting Theorem": Root-clustering in a subregion Ω_2 (3.2) of the complex plane for a system matrix A is equivalent to root-clustering in a shifted subregion $\hat{\Omega}_2$ (3.4) of the complex plane for a "shifted eigenvalue" system matrix \hat{A} , where

$$\hat{\Omega}_2 = \{(\hat{x}, \hat{y}): (\hat{x}/v)^f + (\hat{y}/w)^g - u < 0\} \quad (3.4a)$$

$$\hat{x} = x - \alpha \quad (3.4b)$$

$$\hat{A} = A - \alpha I. \quad (3.4c)$$

Proof. Using (3.4c)

$$\lambda(A + E) = \lambda(A + \alpha I + E) = \lambda(\hat{A} + E) + \alpha$$

and for the real coordinate of the subregion, with (3.4b)

$$\hat{x} = x - \alpha.$$

Thus, the bound on the perturbation E can be computed for the shifted subregion and "shifted eigenvalue" system matrix, being valid for the original subregion and system matrix.

Remark 3.2: Theorem 3.2 can also be proved using the General Lyapunov Equation; that is, when (3.4c) is substituted in the GLE of the shifted subregion (with coefficients (2.4) for this shifted subregion) after some standard manipulations it can be shown that for subregions that can be shifted the solution P of the GLE of the shifted subregion equals the solution P of the GLE of the original subregion.

Remark 3.3: For convenience and motivated by fact that the coefficient c_{10} becomes zero in the bound (see (3.1b) and Fact 4) the subregion is shifted along the real axis to the origin. For some very specific cases, the best bound is not obtained by shifting exactly to the origin. In that case, one can find the "optimal shift" along the real axis which yields the best bound.

Remark 3.4: The result (3.1) is a function of S . In [7] it is shown that for similar cases an optimization (maximization) of $\mu(a(S), b(S))$ reduces conservatism. This can be done with the program fminu.m in Matlab [9]. The power of this method will be illustrated with examples in Section V.

Table I shows the results of this section for some second order subregions. When the regions satisfy (3.2), the parameters holding for the shifted subregions are presented.

TABLE 1

RESULTS FOR DIFFERENT SUBREGIONS. c_{pq} IS FOUND WITH (2.4) AND (2.3). THE EIGENVALUES ARE LOCATED IN THE SUBREGION Ω_1 (2.3) OR Ω_2 (2.4) IN THEOREM 3.1 WITH a AND b AS BELOW IS SATISFIED. THE EQUATIONS AFTER THE '~' ARE OBTAINED FROM THE SHIFTED SUBREGION (THEOREM 3.2)

Subregion	$\Omega_2 = (x,y):$	c_{00}	c_{10}	c_{11}	c_{20}	a	b
Circle ~	$(x/r)^2 + (y/r)^2 - 1 < 0$	-1	0	r^2	0	$ r^2 \ P\ \ S\ ^2$	$2 r^2 \ S\ \ PAS\ $
Parabola ~	$4px + y^2 < 0$	0	$2p$	$\frac{1}{2}$	$-\frac{1}{4}$	$\frac{1}{2} \ P\ \ S\ ^2 + \frac{1}{2} \ S\ \ PS\ $	$4p \ S\ \ PS\ + \ S\ \ PAS\ + \frac{1}{2} \ S^T A^T\ \ PS\ + \frac{1}{2} \ S\ \ A^T PS\ $
ellipse ~	$(x/v)^2 + (y/w)^2 - 1 < 0$	-1	0	$\frac{1}{2}(v^2 + w^2)$	$\frac{1}{4}(v^2 - w^2)$	$\frac{1}{2}(v^2 + w^2) \ P\ \ S\ ^2 + \frac{1}{2}(v^2 - w^2) \ S\ \ PS\ $	$(v^2 + w^2) \ S\ \ PAS\ + \frac{1}{2} v^2 - w^2 (\ S^T A^T\ \ PS\ + \ S\ \ A^T PS\)$
Vertical strip ~	$x^2 - \beta^2 < 0$	$-\beta^2$	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2} \ P\ \ S\ ^2 + \frac{1}{2} \ S\ \ PS\ $	$\ S\ \ PAS\ + \frac{1}{2} \ S^T A^T\ \ PS\ + \frac{1}{2} \ S\ \ A^T PS\ $
Ride quality [6]	1 $-x^2 - y^2 + \rho_1^2 < 0$ 2 $x^2 + y^2 - \rho_2^2 < 0$ 3 $-\omega^2 + y^2 < 0$	ρ_1^2 $-\rho_2^2$ $-\omega^2$	0 0 0	1 1 $\frac{1}{2}$	0 0 $-\frac{1}{4}$	$\ P\ \ S\ ^2$ $\ P\ \ S\ ^2$ $\frac{1}{2} \ P\ \ S\ ^2 + \frac{1}{2} \ S\ \ PS\ $	$2\ S\ \ PAS\ $ $2\ S\ \ PAS\ $ $\ S\ \ PAS\ + \frac{1}{2} \ S^T A^T\ \ PS\ + \frac{1}{2} \ S\ \ A^T PS\ $

IV. PERTURBATION BOUNDS FOR STRUCTURED UNCERTAINTIES

For structured uncertainties it is possible to obtain less conservative results. In the next theorem we obtain a perturbation bound for this case.

Theorem 4.1 The same assumptions as in Theorem 3.1 (1)-(4) are made. The system is given by (2.1) with the perturbation as in (2.2). The eigenvalues of the perturbed system are located in the described subregion (2.3) if

$$|\epsilon| < \mu = \frac{b}{2a} \left(-1 + \sqrt{1 + \frac{4a}{b}} \right) \quad (4.1)$$

where

$$a \equiv \left\| \sum_{i,j} \sum_i |P_{ij}^*| \right\|, \quad b \equiv \left\| \sum_{i=1}^n |P_i^*| \right\|$$

$$P_{ij}^* \equiv S^{-1} P_j S^{-1}, \quad P_i^* \equiv S^{-1} P S^{-1}$$

$$I \equiv c_{11}(F_i^T P E_i + F_j^T P E_i) + c_{20}((F_i E_j)^T P + P E_i E_j + (F_j F_i)^T P + P F_j F_i)$$

$$\equiv c_{10}(E^T P + P E) + c_{11}(F_i^T P A + A^T P F_i) + c_{20}((E_i A)^T P + P E_i A + (A F_i)^T P + P A F_i)$$

where $|A| = \{|a_{ij}|\}$, with $A = \{a_{ij}\}$ a $n \times n$ matrix.

Proof. See Appendix.

Remark 4.1 For first order subregions, (4.1) becomes $\mu = 1/b$ (Remark 3.1).

Remark 4.2 Optimization of $\mu(a(S), b(S))$ in (4.1) reduces conservatism (see Remark 3.4).

Similar to Section III, it is possible to work out this general result for different subregions. This is done in Table II.

V. EXAMPLES FOR DIFFERENT SECOND ORDER SUBREGIONS

Root Clustering in a Circle with Unstructured Uncertainties

The result will now be illustrated with an example. We take the same plant matrix as Yedavalli [5]

$$A = \begin{bmatrix} -4 & 3 & -0.1 \\ 0 & 2 & -3.1 \end{bmatrix} \quad (5.1)$$

with eigenvalues $\lambda_1 = -1.2$, $\lambda_2 = -3.5$. The circular subregion is defined with $\alpha = -1$, $\beta = 1$, $\gamma = 1$, $\delta = 2$ in (3.2). First, taking $Q = S = I$, (3.1) gives the upper perturbation bound

$$\mu_{S=I} = 0.0370$$

where the result in [5] is $\mu_{S=I} = 0.0341$. This is an improvement of 8.5%.

Secondly, we shift the center of circle to the origin (Theorem 3.2) and apply our root clustering Theorem 3.1 for the system matrix A in the subregion Ω_2 (3.4a). This gives

$$A = \begin{bmatrix} -0.3 & -0.1 \\ 0.2 & 0.6 \end{bmatrix}$$

with eigenvalues $\lambda_1(A) = -0.2$, $\lambda_2(A) = 0.5$. The upper bound on $\|E\|$ becomes

$$\mu_{S=I} = 0.3782$$

Comparing the result $\mu_{S=I}$ to $\mu_{S=I}$ [5] we see that the bound is improved by 1009%.

Thirdly, for the shifted subregion Ω_2 and the "shifted eigenvalue" matrix A , optimizing S gives the solutions with (2.5)

$$S = \begin{bmatrix} 0.9013 & 0.0723 \\ 0.0760 & 0.8494 \end{bmatrix}, \quad P = \begin{bmatrix} 0.9098 & 0.2981 \\ 0.2981 & 1.1395 \end{bmatrix}$$

With (3.1) this gives

$$\mu_{S=S_{opt}} = 0.3865$$

Comparing $\mu_{S=S_{opt}}$ and $\mu_{S=I}$ an improvement of 2.2% is achieved.

TABLE II
RESULTS FOR DIFFERENT SUBREGIONS: c_{pq} IS FOUND WITH (2.4) AND (2.3). THE EIGENVALUES ARE LOCATED IN THE SUBREGION Ω_2 IF THE CONDITIONS IN THEOREM 4.1 WITH P_i AND P_i AS BELOW ARE SATISFIED

Subregion	$\Omega_2 = (x, y)$:	c_{00}	c_{10}	c_{11}	c_{20}	P_0	P_1
circle	$(x-\alpha)^2 + y^2 - r^2 < 0$	$\alpha^2 - r^2$	$-\alpha$	1	0	$E_1^T P E_1 + E_1^T P E_1$	$-\alpha(E_1^T P + P E_1) + (E_1^T P A + A^T P E_1)$
parabola	$-4\alpha x + 4px + y^2 < 0$	$-4p\alpha$	$2p$	$\frac{1}{2}$	$-\frac{1}{4}$	$\frac{1}{2}(E_1^T P E_1 + E_1^T P E_1)$ $-\frac{1}{4}(E_1^T E_1^T P + P E_1 E_1 + E_1^T E_1^T P + P E_1 E_1)$	$2p(E_1^T P + P E_1) + \frac{1}{2}(E_1^T P A + A^T P E_1)$ $-\frac{1}{4}(A^T E_1^T P + P E_1 A + E_1^T A^T P + P A E_1)$
ellipse	$((x-\alpha)/v)^2 + (y/w)^2 - 1 < 0$	$(\alpha/v)^2 - 1$	$-\alpha v^{-2}$	$\frac{1}{2}(v^2 + w^2)$	$\frac{1}{4}(v^2 - w^2)$	$\frac{1}{2}(v^2 + w^2)(E_1^T P E_1 + E_1^T P E_1)$ $+\frac{1}{4}(v^2 - w^2)(E_1^T E_1^T P + P E_1 E_1 + E_1^T E_1^T P + P E_1 E_1)$	$-\alpha v^{-2}(E_1^T P + P E_1) + \frac{1}{2}(v^2 + w^2)(E_1^T P A + A^T P E_1)$ $+\frac{1}{4}(v^2 - w^2)(A^T E_1^T P + P E_1 A + E_1^T A^T P + P A E_1)$
Vertical strip	$(x-\alpha)^2 - \beta^2 < 0$	$\alpha^2 - \beta^2$	$-\alpha$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}(E_1^T P E_1 + E_1^T P E_1)$ $+\frac{1}{4}(E_1^T E_1^T P + P E_1 E_1 + E_1^T E_1^T P + P E_1 E_1)$	$-\alpha(E_1^T P + P E_1) + \frac{1}{2}(E_1^T P A + A^T P E_1)$ $+\frac{1}{4}(A^T E_1^T P + P E_1 A + E_1^T A^T P + P A E_1)$
Ride quality [6]	1. $\rho_1^2 < x^2 + y^2$ 2. $x^2 + y^2 < \rho_2^2$ 3. $-w^2 + y^2 < 0$	ρ_1^2 $-\rho_2^2$ $-w^2$	0 0 0	-1 1 $\frac{1}{2}$	0 0 $-\frac{1}{4}$	$-(E_1^T P E_1 + E_1^T P E_1)$ $E_1^T P E_1 + E_1^T P E_1$ $\frac{1}{2}(E_1^T P E_1 + E_1^T P E_1)$ $-\frac{1}{4}(E_1^T E_1^T P + P E_1 E_1 + E_1^T E_1^T P + P E_1 E_1)$	$-(E_1^T P A + A^T P E_1)$ $E_1^T P A + A^T P E_1$ $\frac{1}{2}(E_1^T P A + A^T P E_1)$ $-\frac{1}{4}(A^T E_1^T P + P E_1 A + E_1^T A^T P + P A E_1)$

Root-Clustering in an Ellipse with Structured Uncertainties

Consider the system and perturbation matrices

$$A = \begin{bmatrix} -0.1 & 0.6 \\ -0.4 & 1 \end{bmatrix}, \quad E_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad E_2 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

With $\lambda_1(A) = 0.2$, $\lambda_2(A) = 0.7$. Taking the ellipsoidal subregion with $\alpha = 0.4$, $v = 0.5$, $w = 0.4$, then the eigenvalues λ_1 and λ_2 are located in this subregion. The bound on ϵ_1 and ϵ_2 , μ , can be computed directly with $Q = S = I$ ((4.1) and Table I)

$$\mu_{\epsilon, S=I} = 0.0395.$$

Using the method proposed in [5], taking $Q = I$, we obtain the perturbation bound

$$\mu_{\epsilon, yrd} = 0.0072.$$

Comparing the bounds $\mu_{\epsilon, S=I}$ with $\mu_{\epsilon, yrd}$ we see that our bound gives an improvement of 449%. A still less conservative result can be obtained by optimizing. The optimal S matrix and the solution for P (2.5) become

$$S = \begin{bmatrix} 0.9144 & -0.4853 \\ -0.8012 & 1.0582 \end{bmatrix}, \quad P = \begin{bmatrix} 1.8312 & -1.6732 \\ -1.6732 & 1.8692 \end{bmatrix}.$$

The error bound then becomes

$$\mu_{\epsilon, S=S_{opt}} = 0.1098.$$

Comparing the bounds $\mu_{\epsilon, S=S_{opt}}$ with $\mu_{\epsilon, S=I}$ we see that optimization gives an improvement of 178%. Optimizing the Q matrix and using the method of Yedavalli, the same bound as $\mu_{\epsilon, yrd}$ is obtained.

The results of this example are illustrated in Fig. 1, where the location of the roots in the ellipsoidal subregion of the unit disk, with the maximum perturbation ($\lambda(A \pm \mu(E_1 + E_2))$), are shown.

Regarding this plot, it is important to note that the roots marked with a "x" are, in general, not the outermost points of a root locus determined by varying ϵ from $-\mu$ to μ in $\lambda(A \pm \epsilon(E_1 + E_2))$.

General Remarks

Remark 5.1: In remarks 3.4 and 4.1, comment is given about the optimal S matrix. This S matrix can be obtained with use of

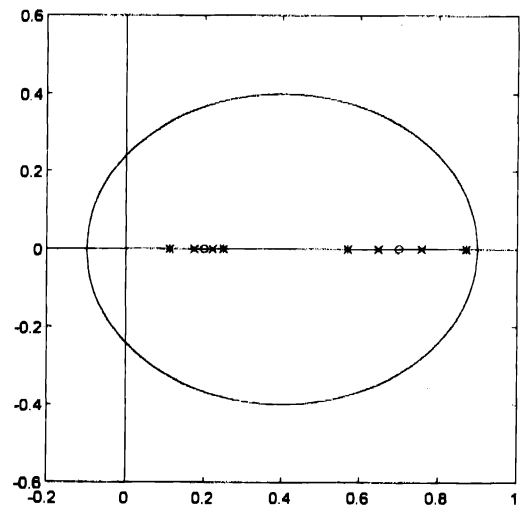


Fig. 1. Root-clustering in the ellipsoidal subregion of the unit disk. The roots of the nonperturbed system are denoted by "o", the roots of the system with the maximum perturbation bound $\mu_{\epsilon, S=I}$ are denoted by "+" and the roots with the bound $\mu_{\epsilon, S=S_{opt}}$ are denoted by "x."

MATLABTM programs for the optimization toolbox [9]. Furthermore, note that the optimal S matrix is not unique. For example, when $\hat{S} = \alpha S$ is substituted, the solution of \hat{P} with (2.5) becomes (where $\hat{Q} = \alpha^2 Q = \alpha^2 S^T S$) $\hat{P} = \alpha^2 P$. Clearly, the bound [(3.1) or (4.1)] obtained for P and S equals the bound obtained for \hat{P} and \hat{S} .

Remark 5.2: It is important to note that the bounds derived are valid for *time-invariant* uncertainties. In the theorems given, however, only the location of the eigenvalues is guaranteed to remain in a specified subregion; this is true for each value of the uncertainty that satisfies the bound. So, when the bound is satisfied, at each time instant, by the *time-varying* uncertainties, then the eigenvalues remain in the subregion. For these time-varying uncertainties, however, this does not guarantee a certain corresponding performance. In engineering practice we are interested in performance; hence we presented bounds that guarantee, beside root-clustering in a subregion, also a corresponding performance.

VI. CONCLUSIONS

In this paper some improved bounds for robust performance of linear systems are obtained. Systems with structured and unstructured uncertainties are treated. For each case, optimization yields better results. For unstructured uncertainties, with shifting of the subregion and the system matrix, a less conservative result is obtained. We have presented our results for various different subregions in the form of tables. With the method proposed it is possible to determine admissible uncertainties of the system to guarantee, for time invariant uncertainties, a certain performance.

APPENDIX

Proof of Theorem 3.1: Recall the Lyapunov root-clustering criterion of Section II. From assumption (2) it follows that there exist a positive definite solution P of the GLE (2.5). If the GLE of the perturbed system matrix $A + E$ has a positive definite "solution" Q_1 for that solution P then the eigenvalues of the perturbed system matrix are located in the specified subregion. Thus, the eigenvalues of the perturbed system are located in Ω_2 if

$$\begin{aligned} c_{00}P + c_{10}((A + E)^T P + P(A + E)) \\ + c_{11}(A + E)^T P(A + E) + c_{20}(((A + E)^T)^2 P \\ + P(A + E)^2) = -Q_1 < 0. \end{aligned}$$

See also Yedavalli [5]. Substitution of (2.5) gives

$$\begin{aligned} &\Leftrightarrow c_{10}(E^T P + PE) + c_{11}(E^T P A + A^T P E + E^T P E) \\ &\quad + c_{20}((A^T E^T + E^T A^T + (E^T)^2)P \\ &\quad + c_{20}P(AE + EA + E^2) - Q < 0 \\ &\Leftrightarrow c_{10}S^{-1}(E^T P + PE)S^{-1} \\ &\quad + c_{11}S^{-1}(E^T P A + A^T P E + E^T P E)S^{-1} \\ &\quad + c_{20}S^{-1}((A^T E^T + E^T A^T + (E^T)^2)PS^{-1} \\ &\quad + c_{20}S^{-1}P(AE + EA + E^2)S^{-1} - S^{-1}S^T S S^{-1} < 0 \\ &\Leftrightarrow c_{10}S^{-1}E^T P S^{-1} + c_{10}S^{-1}P E S^{-1} + c_{11}S^{-1}E^T P A S^{-1} \\ &\quad + c_{11}S^{-1}A^T P E S^{-1} + c_{11}S^{-1}E^T P E S^{-1} \\ &\quad + c_{20}S^{-1}A^T E^T P S^{-1} + c_{20}S^{-1}E^T A^T P S^{-1} \\ &\quad + c_{20}S^{-1}(E^T)^2 P S^{-1} + c_{20}S^{-1}P A E S^{-1} \\ &\quad + c_{20}S^{-1}P E A S^{-1} + c_{20}S^{-1}P E^2 S^{-1} - I < 0. \quad (A.1) \end{aligned}$$

The left side of (A.1) is a Hermitian matrix. A Hermitian matrix has the property that all the eigenvalues (λ_k) are real and that it is negative definite if and only if all its eigenvalues are negative. Further, from [10], the property $\lambda_k(A - I) < 0 \Leftrightarrow \lambda_k(A) - 1 < 0 \Leftrightarrow \|A\| - 1 < 0$ and $\|A\| + \|B\| > \|A + B\|$, with $\|A\| =$ the spectral norm of the matrix A (the largest singular value of A), gives

$$\begin{aligned} &\|c_{10}S^{-1}E^T P S^{-1}\| + \|c_{10}S^{-1}P E S^{-1}\| + \|c_{11}S^{-1}E^T P A S^{-1}\| \\ &\quad + \|c_{11}S^{-1}A^T P E S^{-1}\| + \|c_{11}S^{-1}E^T P E S^{-1}\| \\ &\quad + \|c_{20}S^{-1}A^T E^T P S^{-1}\| + \|c_{20}S^{-1}E^T A^T P S^{-1}\| \\ &\quad + \|c_{20}S^{-1}(E^T)^2 P S^{-1}\| + \|c_{20}S^{-1}P A E S^{-1}\| \\ &\quad + \|c_{20}S^{-1}P E A S^{-1}\| + \|c_{20}S^{-1}P E^2 S^{-1}\| - 1 < 0 \\ &\approx 2\|c_{10}S^{-1}P\|\|E\|\|S^{-1}\| + 2\|c_{11}S^{-1}\|\|E^T\|\|P A S^{-1}\| \\ &\quad + \|c_{11}S^{-1}\|\|E^T\|\|P\|\|E\|\|S^{-1}\| \\ &\quad + 2\|c_{20}S^{-1}A^T\|\|E^T\|\|P S^{-1}\| + 2\|c_{20}S^{-1}\|\|E^T\|\|A^T P S^{-1}\| \\ &\quad + 2\|c_{20}S^{-1}\|\|(E^T)^2\|\|P S^{-1}\| - 1 < 0 \end{aligned}$$

$$\begin{aligned} &\Leftrightarrow (|c_{11}|\|P\|\|S^{-1}\|^2 + 2|c_{20}|\|S^{-1}\|\|P S^{-1}\|)\|E\| \\ &\quad + (2|c_{10}|\|S^{-1}\|\|P S^{-1}\| + 2|c_{11}|\|S^{-1}\|\|P A S^{-1}\| \\ &\quad + 2|c_{20}|\|S^{-1}\|A\|\|P S^{-1}\| \\ &\quad + 2|c_{20}|\|S^{-1}\|\|A^T P S^{-1}\|)\|E\| - 1 < 0 \end{aligned}$$

$$\Leftrightarrow a\|E\|^2 + b\|E\| - 1 < 0 \quad (A.2)$$

with

$$\begin{aligned} a &\equiv |c_{11}|\|P\|\|S^{-1}\|^2 + 2|c_{20}|\|S^{-1}\|\|P S^{-1}\| \\ b &\equiv 2|c_{10}|\|S^{-1}\|\|P S^{-1}\| + 2|c_{11}|\|S^{-1}\|\|P A S^{-1}\| \\ &\quad + 2|c_{20}|\|S^{-1}\|A\|\|P S^{-1}\| + \|S^{-1}\|\|A^T P S^{-1}\| \end{aligned}$$

$$\Leftrightarrow \|E\| < -\frac{b}{2a} + \frac{1}{2a} \cdot \sqrt{b^2 + 4a}.$$

□

Proof of Theorem 4.1: The first part of the proof is the same as the proof of Theorem 3.1. We continue with (A.1)

$$\begin{aligned} &\Leftrightarrow S^{-1}\{c_{10}(E^T P + PE) + c_{11}(E^T P A + A^T P E) \\ &\quad + c_{20}((E A)^T P + (A E)^T P + P(AE + EA)) \\ &\quad + c_{11}E^T P E + c_{20}((E^T)^2 P + P E^T)\}S^{-1} - I < 0. \quad (A.3) \end{aligned}$$

Now the substitution $E \rightarrow \sum_{i=1}^m \epsilon_i E_i$ is made (2.2). Abbreviating

$$P_i^* \equiv S^{-1} P_i S^{-1}; \quad P_{ij}^* \equiv S^{-1} P_{ij} S^{-1}$$

$$\begin{aligned} P_i &\equiv c_{10}(E_i^T P + P E_i) + c_{11}(E_i^T P A + A^T P E_i) \\ &\quad + c_{20}((E_i A)^T P + P E_i A + (A E_i)^T P + P A E_i) \\ P_{ij} &\equiv c_{11}(E_i^T P E_j + E_j^T P E_i) \\ &\quad + c_{20}((E_i E_j)^T P + P E_i E_j + (E_j E_i)^T P + P E_j E_i) \end{aligned}$$

$$\Leftrightarrow \sum_{i=1}^m \epsilon_i P_i^* + \sum_{i=1}^m \sum_{j=1}^m \epsilon_i \epsilon_j P_{ij}^* - I < 0.$$

The left side of this inequality is a Hermitian matrix. From [10], the property

$$\lambda_k(A - I) < 0 \Leftrightarrow \lambda_k(A) - 1 < 0 \Leftrightarrow \|A\| - 1 < 0 \text{ gives}$$

$$\Leftrightarrow \left\| \sum_{i=1}^m \epsilon_i P_i^* + \sum_{i=1}^m \sum_{j=1}^m \epsilon_i \epsilon_j P_{ij}^* \right\| - 1 < 0$$

$$\Leftrightarrow \left\| \sum_{i=1}^m \epsilon_i P_i^* \right\| + \left\| \sum_{i=1}^m \sum_{j=1}^m \epsilon_i \epsilon_j P_{ij}^* \right\| - 1 < 0$$

$$\Leftrightarrow \epsilon \left\| \sum_{i=1}^m |P_i^*| \right\| + \epsilon^2 \left\| \sum_{i=1}^m \sum_{j=1}^m |P_{ij}^*| \right\| - 1 < 0$$

$$\Leftrightarrow a\epsilon^2 + b\epsilon - 1 < 0 \quad (A.4)$$

$$\text{with } a \equiv \left\| \sum_{i=1}^m \sum_{j=1}^m |P_{ij}^*| \right\|, \quad b \equiv \left\| \sum_{i=1}^m |P_i^*| \right\|$$

$$|\epsilon| < -\frac{b}{2a} + \frac{1}{2a} \cdot \sqrt{b^2 + 4a}$$

and where $|A| = \{|a_{ij}|\}$, with $A = \{a_{ij}\}$, a $n \times n$ matrix.

REFERENCES

- [1] A. A. Abdul-Wahab, "Lyapunov bounds for root clustering in the presence of system uncertainty," *Int J Syst Sci*, vol 21, no 12, pp 2603-2611, 1990
- [2] —, "Perturbation bounds for root-clustering of linear discrete-time systems," *Int J Syst Sci*, vol 22, no 10, pp 1775-1783, 1991
- [3] R. K. Yedavalli, "Counter-example to 'Perturbation bounds for root-clustering of linear continuous-time systems,'" *Int J Syst Sci*, vol 23, no 4, pp 661-662, 1992
- [4] W. Bakker and J. S. Luo, "Comment and counter example to 'Perturbation bounds for root-clustering of linear discrete-time systems,'" *Int J Syst Sci*, vol 24, no 11, pp 2205-2206, 1993
- [5] R. K. Yedavalli, "Robust root clustering for linear uncertain systems using generalized Lyapunov theory," *Automatica*, vol 29, no 1, pp 237-240, 1993
- [6] S. Gutman and E. I. Jury, "A general theory for matrix root-clustering in subregions of the complex plane," *IEEE Trans Automat Contr*, vol AC-26, no 4, pp 853-863, 1981
- [7] J. S. Luo, A. Johnson, and P. P. J. van den Bosch, 'New Lyapunov robustness bounds for pole-assignment in a specified region' in *Proc IFAC World Congress*, vol 2, 1993, pp 495-498
- [8] K. Furuta and S. B. Kim, "Pole assignment in a specified disk," *IEEE Trans Automat Contr*, vol AC 32, no 5, pp 423-427, 1987
- [9] A. Grace, *Optimization Toolbox User's Guide for Use with MATLAB*, South Natick, MA: The MathWorks, 1990
- [10] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Orlando, FL: Academic, 1985

Comments on "Strictly Positive Real Transfer Functions Revisited"

H. J. Marquez and C. J. Damaren

Abstract—In the above paper,¹ the distinction between weak and strong strictly positive real (SPR) functions was addressed, and the feedback interconnection of a weak SPR system and a passive one was shown to be stable. The purpose of this note is to show that the proof of this lemma is actually incorrect.

1 INTRODUCTION

The concepts of passivity and strict positive realness have been an important area of research for the last three decades. These investigations have brought a better understanding of these ideas and their applications, but also an ever increasing mismatch in the terminology adopted by different authors. The most widely accepted definitions of passivity and strict passivity are the following [1]. Define a real inner product $\langle \cdot, \cdot \rangle_T$ by

$$\langle \cdot, \cdot \rangle_T = \int_0^T \dot{\cdot}^T(t) \dot{\cdot}(t) dt \quad (1)$$

and let L_{2e}^n be the space of all functions $\dot{\cdot}: R^+ \rightarrow R^n$ which satisfy $\|\dot{\cdot}\|_2^2 = \langle \dot{\cdot}, \dot{\cdot} \rangle_T < \infty$, $\forall T \in R^+$ (R^+ is the set of positive real numbers).

Passivity $H: L_{2e}^n \rightarrow L_{2e}^n$ is said to be passive if there exists $\beta \in R$ such that

$$\langle \dot{\cdot}, H\dot{\cdot} \rangle_T \geq \beta \quad \forall \dot{\cdot} \in L_{2e}^n, \quad \forall T \in R^+ \quad (2)$$

Manuscript received January 24, 1994; revised April 11, 1994.

The authors are with the Department of Engineering, Royal Roads Military College, FMO Victoria, British Columbia V0S 1B0 Canada.
IEEE Log Number 9408546

¹R. Lozano-Leal and S. Joshi, *IEEE Trans Automat Contr*, vol 35, no 11, pp 1243-1245, Nov 1990.

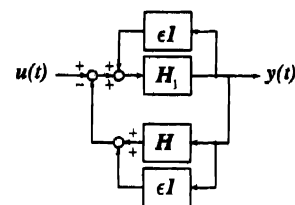


Fig 1 The feedback system S .

Strict Passivity $H: L_{2e}^n \rightarrow L_{2e}^n$ is said to be strictly passive if there exists $\delta > 0$, and $\beta \in R$ such that

$$\langle \dot{\cdot}, H\dot{\cdot} \rangle_T \geq \delta \|\dot{\cdot}\|_2^2 + \beta, \quad \forall \dot{\cdot} \in L_{2e}^n, \quad \forall T \in R^+ \quad (3)$$

For linear systems these definitions are closely related to the concept of strictly positive real (SPR). See the above paper¹ for the definitions of weak and strong SPR. From these definitions, it is straightforward that a linear time-invariant system whose transfer function is (weak or strong) SPR is passive but, in general, not strictly passive. For example the system $H(s) = k/(s+a)$, $k > 0$, and $a > 0$ is SPR (and so passive), however, it is not strictly passive since $\text{Re}[H(j\omega)] \rightarrow 0$ as $\omega \rightarrow \infty$, and therefore no δ can be found to satisfy (3). The main problem with this result is that it renders the passivity theorem nearly inapplicable for linear systems, since only biproper or improper linear systems can be strictly passive.

Motivated by this observation the authors¹ considered the class of systems which satisfy the inequality (6b)¹

$$\frac{\int_0^\infty \dot{y}^T u dt + \beta}{\int_0^\infty \dot{u}^T u dt} > 0 \quad (4)$$

where $\dot{y} = H\dot{u}$, and (4) is valid for all inputs u such that $\|\dot{u}\|_2/\|u\|_2 < \infty$. This class of systems was shown to be equivalent to those which are weak SPR and played a central role in Lemma 1. There, it was claimed that the feedback interconnection of a passive system and one that satisfies inequality (4) (see Fig 1¹) is stable. The intention of this note is to show that the proof of Lemma 1¹ is not valid. We notice here that the condition $\|\dot{u}\|_2/\|u\|_2 < \infty$ is not used in the proof of Lemma 1, and therefore in the remainder of this note it is disregarded.

Using our notation, (4) can be rewritten as

$$\frac{\langle \dot{u}, H\dot{u} \rangle_T + \beta}{\langle \dot{u}, \dot{u} \rangle_T} > 0 \quad (5)$$

A number of comments must be made concerning this definition. In the first place, it is incomplete since for expression (4) to be well defined, it is necessary for the function u to belong to the space L_2^1 . This is an important point. In fact, this issue renders incorrect the proof of Lemma 1¹. Notice that in the proof of Lemma 1¹ it is necessary to consider precisely the case where u is not in L_2 . Yet more appropriate is to use extended spaces and rewrite (4) as

$$\lim_{T \rightarrow \infty} \frac{\langle \dot{u}, H\dot{u} \rangle_T + \beta}{\langle \dot{u}, \dot{u} \rangle_T} > 0, \quad \forall \dot{u} \in L_{2e}^n \quad (6)$$

We now analyze the proof of Lemma 1 (Appendix II¹). Since $u_1 = -\dot{y}$ and $y_1 = u$ (refer to Fig 1¹)

$$\frac{\langle \dot{u}, H\dot{u} \rangle_T + \beta}{\langle \dot{u}, \dot{u} \rangle_T} + \frac{\langle \dot{u}_1, H_1 \dot{u}_1 \rangle_T + \beta_1}{\langle \dot{u}_1, \dot{u}_1 \rangle_T} = \frac{\beta + \beta_1}{\langle \dot{u}, \dot{u} \rangle_T} \quad (7)$$

Here H_1 is passive, and therefore the second term on the left hand side of (7) is greater than or equal to zero, while H satisfies

inequality (6). To show stability, the authors reason by contradiction as follows: assume that $u \notin L_2$, and take limits on both sides of (7) as $T \rightarrow \infty$. In this case, the right-hand side tends toward zero and therefore, the left-hand side also tends toward zero. Hence the authors conclude that there is a contradiction since, by (6), the left-hand side is actually greater than zero. Therefore it must be true that $u \in L_2$.

This reasoning is fallacious, however, unless condition (6) is strengthened by requiring that the following be satisfied

$$\inf_{u \in L_2} \left[\lim_{T \rightarrow \infty} \frac{\langle u, Hu \rangle_T + \beta}{\langle u, u \rangle_T} \right] \geq \delta > 0 \quad (8)$$

In other words, there are only two possibilities of interest in (7)

$$1) \quad \inf_{u \in L_2} \left[\lim_{T \rightarrow \infty} \frac{\langle u, Hu \rangle_T + \beta}{\langle u, u \rangle_T} \right] \geq \delta > 0 \quad (9)$$

If this is the case then indeed there is a contradiction in (7). Condition (9), however, implies that the system is strictly passive and therefore Lemma 1 becomes a restatement of the passivity theorem (see, for example, [1]) i.e., it says nothing about weak SPR functions.

$$2) \quad \inf_{u \in L_2} \left[\lim_{T \rightarrow \infty} \frac{\langle u, Hu \rangle_T + \beta}{\langle u, u \rangle_T} \right] = 0 \quad (10)$$

In this case there is no contradiction in (7) since the left-hand side also tends toward zero for some function u without violating condition (6) (in the same way $1/n^2 \rightarrow 0$ as $n \rightarrow \infty$, for all $p \in R^+$ ≥ 1).

As a final remark we make the following observations, which emphasize the distinction between weak and strong SPR. It is relatively easy to show that the feedback combination of a (possibly nonlinear) passive plant and a strong SPR compensator is stable. The result can be proved by defining the loop transformation shown in Fig. 1 and noting that it does not alter the stability properties of the original system. It is then straightforward to show that for small enough $\epsilon > 0$, the system $H'_1 = (1 - H_1)^{-1}H_1$ is passive while $H' = H + 1$ is strictly passive and therefore stability follows from the passivity theorem.

The case of a weak SPR system is, however, very different as shown in the following example.

Example 1 Consider the linear time-invariant system $H(s) = (s+c)/[(s+a)(s+b)]$ and let $H'(s) = H(s)/[1-\epsilon H(s)]$. We have

$$\begin{aligned} H'(j\omega) + H'(-j\omega) \\ = 2 \frac{(abc - \epsilon c^2) + \omega^2(a+b-\epsilon-\epsilon)}{(ab - \epsilon c - \omega^2)^2 + (a+b-\epsilon-\epsilon)^2} > 0 \end{aligned} \quad (11)$$

if and only if

$$abc - \epsilon c^2 > 0 \quad a+b-\epsilon-\epsilon > 0 \quad (12)$$

if $a+b > \epsilon$, we can always find an $\epsilon > 0$ that satisfies (12). If, however, $a+b = \epsilon$ (i.e., when $H(s)$ is weak SPR), no such $\epsilon > 0$ exists.

II. CONCLUSIONS

The proof of Lemma 1¹ is incorrect. Since this note does not prove that the feedback interconnection of a passive plant and a weak SPR controller is stable, we conclude that it remains an open question.

REFERENCES

- [1] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*. New York: Academic, 1975.

On Interval Polynomials with No Zeros in the Unit Disc

V. Blondel

Abstract—We give a necessary condition for an interval polynomial to have no zeros in the closed unit disc. The condition is expressed in terms of the two first intervals.

The stability analysis of polynomials subject to structured uncertainty has received considerable attention this last decade (see [2] for an historical overview, references related to this contribution include [1], [3], [5], [8], and [9]).

In this note we give a necessary condition for an interval polynomial

$$P = \{a_0 + a_1z + \dots + a_nz^n \mid \underline{a}_i \leq a_i \leq \bar{a}_i\}$$

to be D -stable, i.e., such that all members of P have no roots in the closed unit disc. Our condition is expressed in terms of the two first intervals only.

In a corollary we show that if $\underline{a}_0 < \bar{a}_0/2$ and $\underline{a}_1 < \bar{a}_1/9$ then P cannot be D -stable.

The results presented here are easy consequences of a little known theorem on analytic functions.

Landau's Theorem Assume that the function f is analytic in the open unit disc $|z| < 1$ and that $f(z) \neq 0, 1$ for all $|z| < 1$. Then

$$|f'(0)| \leq 2|f(0)|(|\log|f(0)|| + 4)$$

where 4 is a constant which can be taken equal to 4.4.

For a proof of this theorem (which is sometime referred to as Landau-Carathéodory theorem) see for example Hille [4, p. 221]. The best possible bound for 4 was given in 1981 by Jenkins [6], it is equal to $4\pi/\Gamma(1/4) = 4.37$.

We now prove our theorem.

Theorem Let $P = \{a_0 + a_1z + \dots + a_nz^n \mid \underline{a}_i \leq a_i \leq \bar{a}_i\}$ be an interval D -stable polynomial and assume that $\bar{a}_0 > \underline{a}_0 > 0$. Then

$$|a_1| \leq 2\underline{a}_0 \left(\log^+ \frac{\underline{a}_0}{\bar{a}_0 - \underline{a}_0} + 1 \right)$$

where $\log^+ t = \max(0, \log t)$.

Proof Define $a_i^* \in [\underline{a}_i, \bar{a}_i]$ by $a_0^* = \min(2\underline{a}_0, \bar{a}_0)$ and choose an arbitrary set of coefficients $a_i^* \in [\underline{a}_i, \bar{a}_i]$ ($i = 2, \dots, n$). Consider the polynomial $p(z)$ defined by

$$p(z) = \frac{1}{\underline{a}_0 - a_0^*} (\underline{a}_0 + a_1z + a_2^*z^2 + \dots + a_n^*z^n)$$

It is easy to see that $p(z)$ never takes the value zero or one in the open unit disc. Indeed

$$p(z) = 0 \Leftrightarrow \underline{a}_0 + \bar{a}_1z + a_2^*z^2 + \dots + a_n^*z^n = 0$$

and

$$p(z) = 1 \Leftrightarrow a_0^* + \bar{a}_1z + a_2^*z^2 + \dots + a_n^*z^n = 0$$

Manuscript received November 4, 1993; revised June 25, 1994.

The author is with the Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology (KTH), S-100 44 Stockholm, Sweden.

IEEE Log Number 9408547.

These two polynomials belong to P , hence $p(z) \neq 0, 1$ when $|z| < 1$. Applying Landau's theorem on $p(z)$ we obtain

$$\left| \frac{\bar{a}_1}{a_0 - a_0^*} \right| \leq 2 \left| \frac{a_0}{a_0 - a_0^*} \right| \left(\left| \log \left| \frac{a_0}{a_0 - a_0^*} \right| \right| + 4.4 \right).$$

Since $a_0^* > a_0 > 0$ we get

$$|\bar{a}_1| \leq 2a_0 \left(\left| \log \frac{a_0}{a_0^* - a_0} \right| + 4.4 \right)$$

The coefficient a_0^* is defined by $a_0^* = \min(2a_0, \bar{a}_0)$. If $2a_0 \leq \bar{a}_0$ then $a_0^* = 2a_0$ and

$$|\bar{a}_1| \leq 2a_0 4.4 = 2a_0 \left(\log^+ \frac{a_0}{\bar{a}_0 - a_0} + 4.4 \right)$$

whereas if $2a_0 > \bar{a}_0$ then $a_0^* = \bar{a}_0$ and

$$|\bar{a}_1| \leq 2a_0 \left(\left| \log \frac{a_0}{a_0^* - a_0} \right| + 4.4 \right) = 2a_0 \left(\log^+ \frac{a_0}{\bar{a}_0 - a_0} + 4.4 \right).$$

The theorem is thus proved

Remarks

- 1) A corresponding theorem can be derived for other stability regions. For Schur stability (no roots outside the open unit disc) we obtain a necessary condition for the stability of interval polynomials with uncertainty in the highest order coefficient.
- 2) It is clear from the proof of theorem that, if $p_1(z) = \alpha + a_1 z + a_2 z^2 + \dots + a_n z^n$ and $p_2(z) = \beta + a_1 z + a_2 z^2 + \dots + a_n z^n$ are both D -stable polynomials, then

$$|\alpha| \leq 2|\beta| \left(\left| \log \left| \frac{\alpha}{\beta} \right| \right| + 4.4 \right)$$

This inequality can be used to derive bounds for other structured uncertainties descriptions.

Corollary Let $P = \{\alpha_0 + a_1 z + \dots + a_n z^n \mid \underline{a}_i \leq a_i \leq \bar{a}_i\}$ be an interval polynomial and assume that $0 < 2\underline{a}_0 < \bar{a}_0$ and $\underline{a}_0 < \bar{a}_1$. Then P cannot be D -stable.

Proof Assume by contradiction that P is D -stable. Since $2\underline{a}_0 \leq \bar{a}_0$, the theorem gives $|\bar{a}_1| \leq 2\underline{a}_0 4.4 = 8.8\underline{a}_0$. But this is a contradiction since $\underline{a}_0 < \bar{a}_1$. The result is thus proved.

REFERENCES

- [1] J. E. Ackermann and B. R. Barmish, "Robust Schur stability of a polytope of polynomials," *IEEE Trans Automat Contr*, vol. 33, pp. 984-986, 1988.
- [2] B. R. Barmish and H. I. Kang, "A survey of extreme point results for robustness," *Automatica*, vol. 29, pp. 13-35, 1993.
- [3] N. K. Bose, E. I. Jury, and E. Zeheb, "On robust Hurwitz and Schur stability," *IEEE Trans Automat Contr*, vol. 33, pp. 1166-1168, 1988.
- [4] E. Hille, *Analytic Function Theory*, vol. II. New York: Ginn, 1962.
- [5] C. V. Hollot and A. C. Bartlett, "Some discrete-time counterparts to Kharitonov's stability criterion for uncertain systems," *IEEE Trans Automat Contr*, vol. 31, pp. 355-356, 1986.
- [6] J. Jenkins, "On explicit bounds in Landau's theorem II," *Canadian J Math*, vol. 33, pp. 550-562, 1981.
- [7] V. L. Kharitonov, "Asymptotic stability of an equilibrium position of a family of systems of linear differential equations," *Differentsialnye Uravneniya*, vol. 14, pp. 1483-1485, 1978.
- [8] A. Rantzer, "Kharitonov's weak theorem holds if and only if the stability region and its reciprocal are convex," *Int J Nonlinear Robust Contr*, to appear.
- [9] C. B. Soh, C. S. Berger, and K. P. Dabke, "A general stability theorem for linear discrete-time systems," *IEEE Trans Automat Contr*, vol. AC-30, pp. 505-507, 1985.

The Logical Control of an Elevator

Derek N. Dyck and Peter E. Caines

Abstract—This paper presents a detailed example of the design of a logical feedback controller for finite state machines. In this approach, the control objectives and associated control actions are formulated as a set of axioms each of the form X implies Y , where X asserts that i) the current state satisfies a set of conditions and ii) the control action y will steer the current state towards a given target state; Y asserts that the next control input will take the value y . An automatic theorem prover establishes which of the assertions X is true, and then the corresponding control y is applied. The main advantages of this system are its flexibility (changing the control law is accomplished through changing only the axioms) and the fact that, by the design of the system, control actions will provably achieve the control objectives. The illustrative design problem presented in this paper is that of the logical specification and logical feedback control of an elevator.

I. INTRODUCTION

The COCOLOG system (from Conditional Observer and Controller LOGic) [1], [2] is a logical system for the state estimation and control of finite state machines. In this approach, the control objectives are formulated as axioms (i.e., necessarily true logical formulas) which relate the current state to a target state. The axioms are each of the form X implies Y , where X asserts that i) the current state satisfies a set of conditions and ii) the control action y will steer the current state towards a given target state; Y asserts that the next control input will take the value y . An automatic theorem prover establishes which of the assertions X is true, and then the corresponding control y is applied. The main advantages of this system are its flexibility (changing the control law is accomplished through changing only the axioms) and the fact that, by the design of the system, control actions will provably achieve the control objectives.

This paper applies the COCOLOG system to an idealized version of an elevator control problem. Section II describes the state, dynamics and control of the elevator. Sections III-V briefly outline the COCOLOG system and present a new logical framework specific to the control of an elevator. Sections VI and VII present the results of computer simulations using an Automatic Theorem Prover (due to Mackling, see [3]) and the conclusions which can be drawn from these results.

II. THE ELEVATOR CONTROL EXAMPLE

The logical control of an elevator is a good example with which to illustrate the operation of the COCOLOG system because it shares with many other discrete event systems the features of i) simplicity of dynamics, ii) combinatorial complexity of state description, and iii) great variety of possible control strategies and resulting trajectories. This section describes the basic set-up of the elevator control problem.

A. The State

The elevator control problem studied in this paper consists of a single elevator in a building with five floors, numbered zero to four. The elevator can handle up to three demands, with the

Manuscript received March 29, 1994. This work was supported in part by NSERC Grant GP001329 and NCE IRIS Project B5.

The authors are with the Department of Electrical Engineering, McGill University, Montréal, Québec, Canada H3A 2A7.
IEEE Log Number 9408548.

TABLE I
ELEVATOR STATE VARIABLES AND THEIR VALUES FOR THE INITIAL STATE OF THE EXAMPLE TRAJECTORY

Current Floor = 2		
Location ₀ = 1	Location ₁ = 2	Location ₂ = 3
Destination ₀ = 2	Destination ₁ = 0	Destination ₂ = 4
Frustration ₀ = 3	Frustration ₁ = 2	Frustration ₂ = 1

demands numbered zero to two. A demand is characterized by the demand state which is given by a triple consisting of a location, a declared destination, and a frustration level. To simplify the state representation, a frustration level of zero is used to indicate an inactive demand. Table I contains the complete set of state variables and their values for the initial state of the example to be presented later.

2. System Dynamics

The dynamics of the elevator are quite simple. At each time step, the elevator either moves up or down one floor, or stays at the same floor, according to the control input. This means the elevator effectively stops at every floor it passes. New arrivals are treated as disturbance inputs: when a person enters the world of the elevator, he or she immediately announces a destination and starts with a frustration of one. The frustration is then incremented at each time step until the individual reaches his or her destination. It is assumed that there are never more demands arriving than the maximum number of demands allowed by the system (if there are, these demands are simply ignored until an active demand is satisfied).

Any person at the same floor as the elevator moves with the elevator (that is, the elevator takes on everybody at each floor it arrives at, even if they are traveling in the opposite direction). When a person's location becomes equal to his or her destination, that person immediately leaves the world of the elevator. That is, when the location and destination of a demand become equal, the frustration of the demand is set to zero, indicating an inactive demand.

The example presented in this paper involves satisfying the active demands in the sense that the controlled trajectory of the elevator results in a state in which the location and destination of each active demand are equal, and the elevator is at rest. This formulation is easily extended to the case where new demands occur at random instants and enter the system if fewer than three demands are active at that instant (see, e.g., [3]).

C. Control Law

The control problem for an elevator may be viewed as a disturbance rejection control problem. The basic objective is to empty the elevator, though other goals may temporarily or permanently override this one. The basic control law must therefore guarantee that in the absence of disturbances and overrides the elevator empties in finite time. Our control strategy for the basic control problem is that of prioritizing persons with a high frustration level.

The simple control law which satisfies these objectives is the control law "Max first" which always attends to the person with the highest frustration level. That is, at each time step, the person with the highest frustration is selected and referred to as Max. If Max is not in the elevator, the elevator moves up or down in the direction of Max's location. If Max is in the elevator, the elevator moves in the direction of Max's destination. Since the dynamics guarantee that Max will be selected when the elevator reaches Max's floor, this control law

satisfies the basic objective of emptying the elevator in finite time in the absence of fresh inputs (i.e., new demands or overrides).

To illustrate the basic COCOLOG property of flexibility with respect to control objectives, two override control objectives are added. In the first, the fourth floor of the building is designated as a hospital, and it is assumed that any person whose destination is the fourth floor is an emergency case who takes priority over all other persons in the world of the elevator. In the second, the objective is changed altogether in case of a fire: as is common practice, when the fire alarm is triggered, the elevator stops at the nearest floor (so that people can take the stairs to safety). In our control strategy the fire objective overrides the hospital emergency objective.

The ease with which these control objectives can be added to the basic control law illustrates the power of the COCOLOG approach to control. It should be emphasized that the "Max first" basic control law is not designed to be an optimal one, but rather one which is simple enough to clearly illustrate its implementation using COCOLOG. It is to be noted that logic control laws of the type: "serve the most frustrated customer first" treat large sets of states as equivalent, since the states of the demands not being served are ignored. Consequently, the logic control law described here succeeds in handling a system in which the number of possible states is estimated at 10 million.

III. COCOLOG

The COCOLOG system is a family of first-order logical theories. The theories provide a logical system for describing and reasoning about the state estimation and control of a given finite input-state-output machine. The COCOLOG system consists of a family of theories because at each time step a new theory is generated, since some axioms are used to describe the inputs and observed outputs of the system (or the next state in the case of complete state observation). More formally, at each discrete time instant k , the k th first order logical theory Th_k is used to determine the next control input $U(k)$ through a set of Conditional Control Axioms (CCA's) with Th_k . The CCA's imply certain values for the control input when certain conditions C_k^m are fulfilled. In logical syntax, the CCA's are of the form

$$C_k^m \rightarrow (U(k) = u^m).$$

This formula, in English, would read: "If formula C_k^m in theory k is true, then this implies that the control input at time step k is equal to the constant u^m ."

In COCOLOG, the control input $U(k)$ is assigned a particular value u^m if a proof exists, in Th_k , of the predicate $U(k) = u^m$. Since each CCA_k^m is an axiom, a sufficient condition for this to hold is the existence of a proof of C_k^m in Th_k . We observe that for finite machines in general, and the elevator example in particular, a finite model (in the technical sense of model theory, see [2] and [7]) for any theory is readily constructed. This establishes the consistency of COCOLOG theories in general and ELCOLOG (see below) in particular.

A. Markovian Fragment

The family of theories in a COCOLOG all have in common a subset of the axioms which are the same in all theories; these axioms are, by definition, the axioms in Th_0 . The transition from theory Th_{k-1} to Th_k is governed by meta-level rules of inference, which cannot be represented within these theories. In the general case, each theory Th_k follows from Th_{k-1} by the addition of axioms, for example the axiom which asserts the equality of the output at time instant k to the observed output value. In practice, however, it is very expensive (in terms of proof time on an automatic theorem prover) to keep increasing the number of axioms as the system evolves over time. An alternative is to restrict the set of axioms to a moving window which only includes the most recent additions to the axioms in Th_0 . This is a so called Markovian fragment of COCOLOG theory, which is described in more detail in [4]. It can be shown that, provided certain conditions are satisfied, the Markovian fragment has the same power to make control decisions as the original COCOLOG (this is the principal mathematical result of [4]).

The rest of this section focuses on the application of COCOLOG to the special case of the elevator control example. More details of the operation of COCOLOG in the general case can be found in [2], [5], and [6].

B. Markovian Fragments for the Elevator Control Example

In the case of the elevator control example, the Markovian fragment is implemented by effectively starting from a new initial state at each time step. Each theory Th_k therefore consists only of the axioms in Th_0 together with a set of axioms which describe the state at time instant k . In the general case, the state description axioms would in fact describe the current state estimate (CSE) set of possible current states which satisfy the predicate CSE_k . However, since the elevator system has full state output, i.e., complete state observation, the current state is known exactly, and hence CSE_k is satisfied only by a single state value.

Our axiomatic formulation of the elevator system does not contain a complete axiomatization of the system state transition function. Although this approach reduces the expressibility of our syntax, this approach has several advantages at both the logical and the implementation level:

- 1) A complete logical model of the system is not used in the elevator example to reason about the behavior of the controlled elevator. This reduction of the axioms to a subset of the Markovian fragment increases the efficiency of the logic controller. We note that the actual (computed) state transition of the system (occurring at each control action) is realized outside the ATP program. This is consistent with our ontological assumption that the controlled dynamical system is an entity distinct from the logic controller and the theories it carries.
- 2) Whenever the system undergoes exogenous disturbance inputs (in our case new passengers and their destinations) the resulting state value is efficiently accepted into the flow of COCOLOG theories. This is because the complete state observation hypothesis implies the new state value is immediately available to the COCOLOG controller.
- 3) Since the current state is treated as an initial state within each theory, the time index can be dropped altogether within each theory. (This would not be possible, however, if a model of the dynamics of the system is used to predict the state of the system at future time instants.) This results in a significant simplification of the conditional control axiom set.

The logic resulting from our restrictions of a complete Markovian fragment specification is sufficiently different from COCOLOG that

it warrants the distinct name ELCOLOG (for ELevator COntroller LOGic). The next section presents the language of ELCOLOG.

IV. THE ELEVATOR CONTROLLER LOGIC ELCOLOG

ELCOLOG is a first order logic which is used in elevator control to prove or disprove control formulas, given a specified state. Associated with any logic is a syntax and a semantic interpretation. The syntax is a set of symbols and the rules for formulating formulas. The semantics is the meaning attached to these symbols.

A. The Language of ELCOLOG

The ELCOLOG language consists of constant symbols, variable symbols, function symbols, and predicate symbols, as well as the punctuation symbols "(", ")", ",", and ".". The precise syntactic rules governing formulas can be found in any logic text, e.g., [7]. This section gives an informal description of the syntax, and a semantic interpretation.

Constant Symbol Set:

The symbols representing zero and the first four positive integers:
0 1 2 3 4

The Skolem constants (used in place of existentially quantified variables): **Vic** **Max**

Vic is the index to the demand of the person whose destination is the hospital, in the case of an emergency (i.e., the victim), and **Max** is the index to the demand of the person with the maximum frustration level.

Variable Symbol Set:

General variables: **a b c**

Variables used for indexing the location, destination, and frustration: **m n**

The indexing variables may only assume values in the constant symbol set: {0, 1, 2, **Vic**, **Max**}. Normally, this restriction would imply the use of types to constrain the semantics and to increase the efficiency of the ATP. However, in our simple example this is instead accomplished through the use of domain predicates, which are added to the relevant logical axioms to restrict their application to instantiations obeying this rule.

Function Symbol Set:

The following are the state projection functions. They are all implicit functions of the time index k . Note that arg is an index and must be one of {**Vic**, **Max**, 0, 1, 2, **m**, **n**}.

- Pn** - the position of the elevator (at the current time instant).
- Pl(arg)** - the location of the person associated with demand #arg.
- Pd(arg)** - the destination of the person associated with demand #arg.
- Pf(arg)** - the frustration of the person associated with demand #arg.

And the input at this time step is also an implicit function of the time index k :

- U** - the control input.

Predicate Symbol Set:

Logical operators:

- $\neg \text{arg}$ - True iff arg is not true.
- $\text{arg}_1 \wedge \text{arg}_2$ - True iff both arg_1 and arg_2 are true.
- $\text{arg}_1 \vee \text{arg}_2$ - True iff either arg_1 or arg_2 is true.
- $\text{arg}_1 \rightarrow \text{arg}_2$ - True iff arg_2 is true whenever arg_1 is true

Arithmetic relations:

- $\text{arg}_1 = \text{arg}_2$ - True iff arg_1 and arg_2 are equal.
- $\text{arg}_1 < \text{arg}_2$ - True iff arg_1 is strictly less than arg_2 .

Predicates, with the state as an implicit argument, used in the conditional control axioms:

- Fire** - True iff the fire alarm is on.
- Emergency** - True iff someone's destination is the fourth floor (the hospital).

To reduce the clutter of parentheses, the following operator precedence is used, listed from highest to lowest: $=$, $<$, \neg , \wedge , \vee , \rightarrow .

B Rules of Inference

Rules of inference determine how new formulas can be derived from a set of formulas (some of which will be axioms) while preserving truth in the process (that is, if all the formulas in the set are true, then the new formula is also true). Formally, for ELCOLOG to be a logic, it is required that the rules of inference be defined so that proofs of formulas can be constructed. For this formal requirement, the two rules of inference used in ELCOLOG are Modus Ponens and the Generalization rule. Stated in English, Modus Ponens is simply: "If A is true, and A implies B is true, then B is true," where A and B are formulas. The Generalization rule is: "If A is true, then A is true for all v ," where v is a variable, and A is a formula in which v is free (i.e., is not governed by a universal or existential quantifier). Also, along with these two rules, several Logical Axiom Schemata are also defined. In these details ELCOLOG is identical to COCOLOG (see [2]), so they will not be repeated here.

V. AXIOMATIC THEORY OF ELCOLOG

As in COCOLOG, the axioms in Th_0 include all the axioms except the axioms describing the current state. The complete set of axioms in Th_k is given in Appendix I. The axioms in Th_0 divide naturally into number relation axioms, control axioms, and override control axioms. The set of axioms in Th_k is formed by adding the current state axioms.

A Number Relation Axioms

The number relation axioms define the domain of positive integers, and also define the properties of the equality and inequality predicates. To make automatic theorem proving tractable, the arithmetic axioms also include several theorems. In particular, included as "oracle information" are the properties of the "less than" predicate ($<$) for all pairs of integers in the domain.

B Control Axioms

The control axioms are built around the simple control law, described in Section II-C, which attends to the person with the highest destination level. These axioms are in the form of Conditional Control Axioms in COCOLOG

$$C^m \rightarrow U = u^m.$$

- first of these is the axiom which expresses the law
- put if all demands are inactive:

$$Pf(\text{Max}) = 0 \rightarrow U = 0.$$

- other axioms in this form are the following.

Go down if Max is waiting at a floor below the elevator:

$$\neg(Pf(\text{Max}) = 0) \wedge Pl(\text{Max}) < Pn \rightarrow U = 1$$

Go down if Max is in the elevator and wanting to go down:

$$\neg(Pf(\text{Max}) = 0) \wedge Pl(\text{Max}) = Pn \wedge Pd(\text{Max}) < Pn \rightarrow U = 1$$

Go up if Max is waiting at a floor above the elevator:

$$\neg(Pf(\text{Max}) = 0) \wedge Pn < Pl(\text{Max}) \rightarrow U = 2. \text{ and}$$

Go up if Max is in the elevator and wanting to go up:

$$\neg(Pf(\text{Max}) = 0) \wedge Pl(\text{Max}) = Pn \wedge Pn < Pd(\text{Max}) \rightarrow U = 2.$$

In addition to these axioms, two other axioms are required to uniquely determine the value of the Skolem constant **Max**. They are

$$Pf(n) \leq Pf(\text{Max})$$

and

$$Pf(n) = Pf(\text{Max}) \rightarrow \neg(n < \text{Max}).$$

The next section discusses the override control axioms and explains how they are added to these basic control axioms.

C. Override Control Axioms

The full power of ELCOLOG, as for COCOLOG, lies in the ease with which it allows control objectives to be modified. In this case, the two override control objectives each use a predicate symbol to indicate whether the override is in effect or not. These predicate symbols are then logically OR'ed into all the basic conditional control axioms. For example, the control axiom in the case where the elevator is empty is: stay put if all demands are inactive; this becomes

$$\text{Fire} \vee \text{Emergency} \vee (Pf(\text{Max}) = 0) \rightarrow U = 0.$$

So in each case the value of the override predicate must be determined, either through the current state axioms in the case of a fire, or through additional axioms in Th_0 , as in the hospital override case. The axiom in this second case, for the hospital override, is

$$\neg(Pf(n) = 0) \wedge Pd(n) = 4 \rightarrow \text{Emergency}.$$

The complete set of axioms, including the hospital override axioms, is given in Appendix I.

D. Current State Axioms

The current state axioms are added to Th_0 to form Th_k . These axioms are very simple, as they consist only of a set of equality predicates which equates each state projection function to the observed output value (recall this system has full state output, i.e., complete state observation). For example, one axiom in this set for the state shown in Fig. 1 is

$$Pl(1) = 2$$

which simply asserts that the location of the person associated with the demand #1 is the second floor. The current state axioms corresponding to Fig. 1 are given in Appendix I.

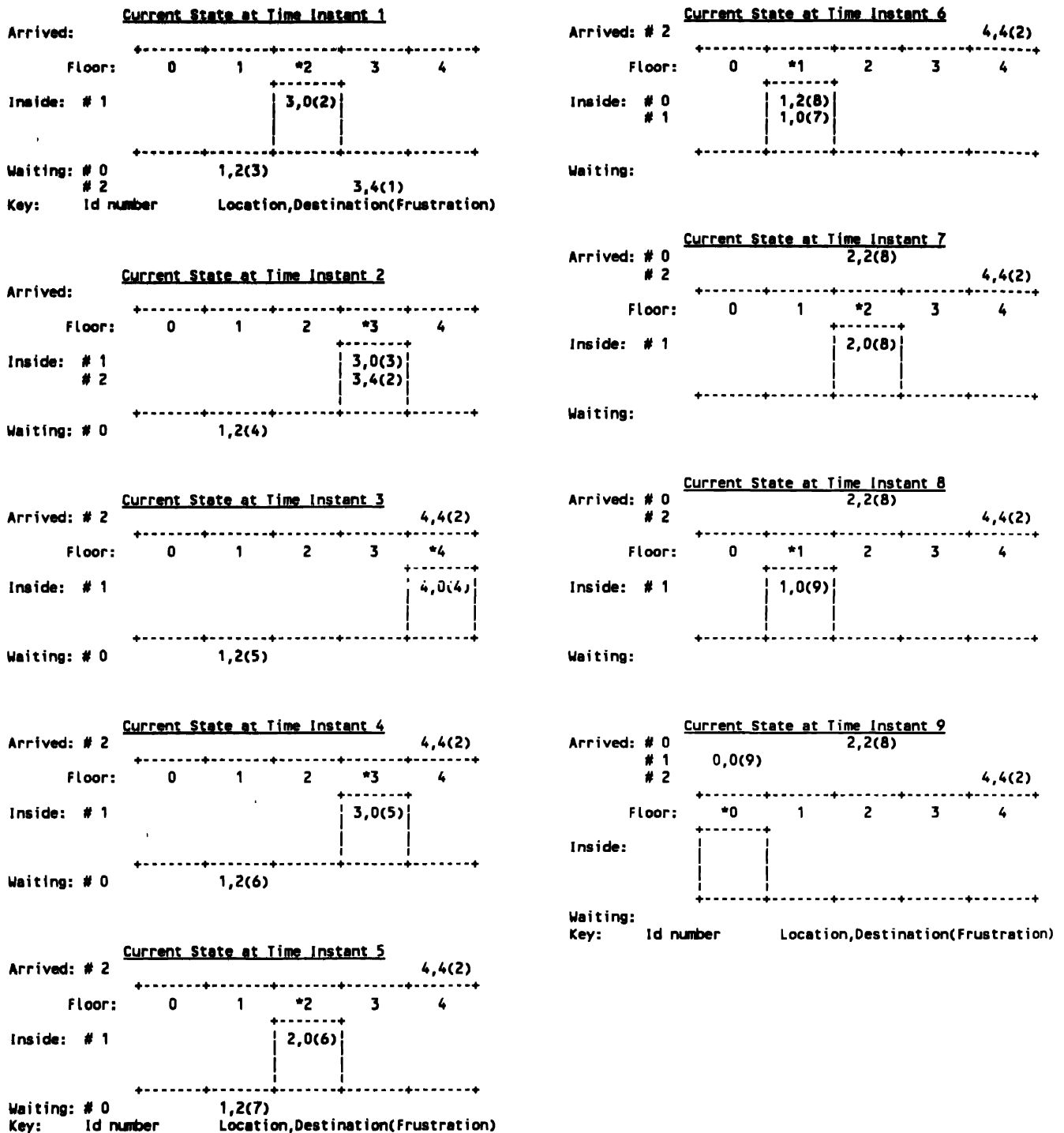


Fig. 1. Elevator control example—hospital override.

VI. IMPLEMENTATION

A. The Automatic Theorem Prover

To determine the control input given an initial state an Automatic Theorem Prover (ATP) is used. This section briefly describes the "Blitzensturm" resolution-refutation theorem prover, developed by Mackling [3], which was used for the simulations. In the propositional logic case (i.e., no variables) resolution is an inference rule which is a generalization of Modus Ponens. This inference rule applies to formulas in canonical clausal form: the formula consists of terms (or the negation of terms) which are OR'ed together. Resolution

works by first searching for two formulas which share a term that appears negated in one formula but not in the other. The new formula is derived by deleting the shared term from both formulas and combining them with the OR operator. For example: the fire override control axiom: $\text{Fire} \rightarrow U = 0$ can be re-written in canonical clausal form as: $\neg \text{Fire} \vee U = 0$. If the current state axiom set includes the formula Fire , then by applying the resolution rule, the new formula $U = 0$ can be derived. In the first-order logic case, the resolution rule involves the process of unification (see [9]).

A proof by refutation starts with a consistent set of axioms, to which is added the negation of the formula to be proven. If

contradiction can be derived using the rules of inference, then at least one of the original formulas must be false, since inference rules cannot derive false formulas from true ones. Since the original set of axioms is consistent, it is the negation of the formula to be proven which must be false, and therefore the formula must be true. For example, the formula to be proven might be $U = 0$, which asserts that the control input is STAY. To prove this, the negation of this formula $\neg(U = 0)$ is added to Th_k . In the case where one of the current state axioms is **Fire**, $U = 0$ can be derived by resolution, as shown above; then $U = 0$ with $\neg(U = 0)$ yields by resolution the empty formula, i.e., a contradiction. The steps leading to the empty formula constitutes a proof of the formula $U = 0$.

A resolution-refutation ATP, defined in terms of these concepts, is an algorithm which recursively applies the resolution inference rule to a set of axioms together with the negation of a formula in order to find the empty formula. In the case where the original formula is, in fact, false, the theorem prover will fail to terminate, since a contradiction cannot be derived from a set of true formulas. Therefore, in practice, it is necessary to run the ATP on all possible control actions simultaneously, and terminate them all as soon as one finds a refutation proof.

We say formula A subsumes formula B if $A \rightarrow B$. For example, $C_1 \vee D$ subsumes $C_1 \vee C_2 \vee D$. It is a fact that, if A subsumes B , then any formula which can be derived by recursively applying the resolution inference rule to A , B , and S can also be derived from only A and S , where S is a given set of additional formulas. The "Blitzensturm" ATP extensively applies the subsumption rule which, in its simplest form, deletes subsumed formulas from the list of formulas at any stage in the resolution process (see [9]). For example, the current state axiom $\neg\text{Fire}$ subsumes the fire override control axiom: $\text{Fire} \rightarrow U = 0$. In this case, the fire override control axiom can be deleted by the ATP, since it is not necessary to prove any control actions. Subsumption is especially useful in keeping the size of the set of formulas manageable while searching for a proof. In particular, note that the combined effect of the resolution rule and the subsumption rule in the above examples is to delete the irrelevant axioms from Th_k depending on whether **Fire** or $\neg\text{Fire}$ appears among the current state axioms. This applies to override controls in general, and so the addition of override controls does not necessarily imply longer proof times.

The ATP has no built-in provision for handling objects of different types. Specifically, the restrictions imposed on indexes are not obeyed by the ATP. However, as explained in [8], the set of axioms can be converted into a single typed language by using domain predicates, which, in ELCOLOG, is the inequality predicate. For example, the first axiom to determine **Max**, in which n is universally quantified, becomes

$$n < 3 \rightarrow \neg(\text{Pf}(\text{Max}) < \text{Pf}(n)).$$

These domain predicates are included in the input file for the ATP.

The Trajectory

The initial state is as shown in Table I. Note that there is one person whose destination is the fourth floor. Appendix II shows the movement of the elevator with the hospital override in effect. In this graphical representation of the state, the elevator travels horizontally, and the triplets of integers display each person's location, destination, and frustration. This trajectory was computed using the Blitzensturm ATP of Mackling [3].

VII. DISCUSSION AND CONCLUSION

A. Computational Complexity

At the time this paper was written, the time required for a single proof was of the order of 100 seconds on a SUN Sparc workstation. This is clearly not suitable for the real time control of an actual elevator. However, significant improvements in proof times are continually being made for T. Mackling's Blitzensturm automatic theorem prover (see [3]), and in view of predicted advances in computer processing speeds it is not unreasonable to expect real time performance for systems of significant complexity in the future.

With regard to more complex systems, it is a fact that the computational complexity of refutation proofs is exponential in the number of axioms [10]. This is the central motivation for the theory of Markovian fragments of COCOLOG [4] (which avoids increasing the size of the axiom set over time). An additional measure to combat increasing complexity is that of a hierarchical control strategy. In this approach, the states of complex systems are aggregated into the state sets of simpler systems at a higher hierarchical level. The computation of control actions will be carried out in a manner that respects this hierarchy.

B. Conclusion

In this paper the elevator control problem is employed to illustrate and apply the underlying concepts of COCOLOG. The essential features of this approach are i) the ability to formulate descriptions of dynamical systems and their control objectives by the use of logical axioms and (ii) a flexibility which permits the easy modification of such axioms, in particular the control axioms. In the elevator example this modification was achieved by the addition of override axioms. In our view, the ease with which we could write in and computationally implement the "emergency" and "fire" overrides to the basic "max first" ELCOLOG control objectives provides evidence for the flexibility of COCOLOG systems, and we suggest the term "objective adaptability" for this property. It should also be emphasized that the use of axioms to formulate control objectives enables the resulting control system to be engineered so as to be provably correct. In general, this property is viewed as being difficult to obtain for finite state machines; for example, in current practise considerable time is spent developing and testing elevator controls.

Finally, we shall mention that experiments are in progress with different COCOLOG formulations of the elevator problem which include simulations of exogenous destination demand flows and permit cyclical ATP iterations which attempt to prove different conditional control theorems.

APPENDIX I

COMPLETE AXIOM SET FOR THE ATP

Equality Axioms

Reflexivity: $a = a$, symmetry: $a = b \rightarrow b = a$, and transitivity: $a = b \wedge b = c \rightarrow a = c$

In addition to these axioms, the appropriate substitution axiom is required for every function and predicate symbol.

Strict Inequality Axioms

Nonreflexivity: $\neg(a < a)$ and transitivity: $a < b \wedge b < c \rightarrow a < c$
Substitution axiom for inequality: $a = b \wedge b < c \rightarrow a < c$

Domain Characterization Axioms

The entire domain consists of positive integers: $a \geq 0$

The entire domain is ordered: $a = b \vee a < b \vee a > b$

Axioms defining the relation between zero and the four positive integers: $0 < 1, 1 < 2, 2 < 3, 3 < 4$

Theorems to aid the Automatic Theorem Prover (oracle information):

$$0 < 2, 0 < 3, 0 < 4, 1 \geq 0, 1 < 3, 1 < 4, 2 \geq 0, 2 \geq 1,$$

$$2 < 4, 3 \geq 0, 3 \geq 1, 3 \geq 2, 4 \geq 0, 4 \geq 1, 4 \geq 2, 4 \geq 3$$

State Projection Functions

P_n is the number of the current floor.

$Pl(n)$ is the location of the person associated with demand $\#n$.

$Pd(n)$ is the destination of the person associated with demand $\#n$.

$Pf(n)$ is the frustration of the person associated with demand $\#n$.

Substitution axioms for equality of the state projection functions

$$m = n \wedge Pl(m) = a \rightarrow Pl(n) = a,$$

$$m = n \wedge Pd(m) = a \rightarrow Pd(n) = a,$$

$$m = n \wedge Pf(m) = a \rightarrow Pf(n) = a$$

Substitution theorems for inequality of the state projection functions

$$m = n \wedge Pl(m) < a \rightarrow Pl(n) < a,$$

$$m = n \wedge Pd(m) < a \rightarrow Pd(n) < a,$$

$$m = n \wedge Pf(m) < a \rightarrow Pf(n) < a$$

Axioms for Control

Fire Override: If there is a fire, stay at the current floor:

$$\text{Fire} \rightarrow U = 0$$

Fire is OR'ed into all of the following control axioms, for clarity, however, it will not appear in the following.

Hospital Override: If someone's destination is the fourth floor (the hospital) then that person has priority

$$n < 3 \wedge \neg(Pf(n) = 0) \wedge Pd(n) = 4 \rightarrow \text{Emergency}.$$

If there is an Emergency, then the following axioms determine Vic

Vic is (an index to) a person:

$$\text{Emergency} \rightarrow (Vic = 0 \vee Vic = 1 \vee Vic = 2)$$

Vic's destination is the fourth floor: $\text{Emergency} \rightarrow Pd(Vic) = 4$

Vic's frustration is not zero: $\text{Emergency} \rightarrow \neg(Pf(Vic) = 0)$

If there is more than one person whose destination is the fourth floor, then the one with the lowest index is treated first

$$n < 3 \wedge \neg(Pf(n) = 0) \wedge Pd(n) = 4 \rightarrow \neg(n < Vic).$$

Control in case of emergency: deliver victim to the fourth floor

$$Pl(Vic) < P_n \rightarrow U = 1, \neg(Pl(Vic) < P_n) \rightarrow U = 2$$

(since Vic will exit on arrival, this last axiom will not cause an attempt to go up from the fourth floor)

Basic Control: The basic control is: attend to the person with maximum frustration level. This is the control action to be taken in the absence of a Fire or an Emergency. To specify this, each of the axioms A_i in the basic control is to be interpreted as: $\neg\text{Fire} \wedge \neg\text{Emergency} \rightarrow A_i$. For clarity and because of space limitations, only the nonoverridden axioms A_i are specified in the following.

Max is (an index to) a person: $\text{Max} = 0 \vee \text{Max} = 1 \vee \text{Max} = 2$
Axioms which uniquely determine Max:

a) No one is more frustrated than Max: $\neg(Pf(n) > Pf(\text{Max}))$

b) Of those who are equally frustrated, none has a lower demand number than Max

$$Pf(n) = Pf(\text{Max}) \rightarrow \neg(n < \text{Max}).$$

Stay put if all demands are inactive: $Pf(\text{Max}) = 0 \rightarrow U = 0$.

Go down if Max is waiting at a floor below the elevator, or if Max is in the elevator and wants to go down

$$Pf(\text{Max}) \neq 0 \wedge Pl(\text{Max}) < P_n \rightarrow U = 1$$

$$Pf(\text{Max}) \neq 0 \wedge Pl(\text{Max}) = P_n \wedge Pd(\text{Max}) < P_n \rightarrow U = 1.$$

Go up if Max is waiting at a floor above the elevator, or if Max is in the elevator and wants to go up.

$$Pf(\text{Max}) \neq 0 \wedge Pl(\text{Max}) > P_n \rightarrow U = 2$$

$$Pf(\text{Max}) \neq 0 \wedge Pl(\text{Max}) = P_n \wedge Pd(\text{Max}) > P_n \rightarrow U = 2.$$

Axioms Describing the Current State

Current state at time instant $k = 1$ is as in Fig. 1

$$\neg\text{Fire}, P_n = 2, Pl(0) = 1, Pd(0) = 2, Pf(0) = 3, Pl(1) = 2,$$

$$Pd(1) = 0, Pf(1) = 2, Pl(2) = 3, Pd(2) = 4, Pf(2) = 1.$$

With the Hospital Override in effect, the next control input should be up ($U = 2$). This is proved by a refutation proof (a contradiction is derived from the negation of the formula to be proved $\neg(U = 2)$).

ACKNOWLEDGMENT

The authors would like to thank T. Mackling and Y. Wei for many insightful and illuminating discussions and to acknowledge the original challenge by W. M. Wonham to formulate the elevator problem in COCOLOG.

REFERENCES

- [1] S. Wang, "Classical and logic based control theory for finite state machines," Ph.D. dissertation, McGill University, Montréal, Québec, 1991.
- [2] P. E. Caines and S. Wang, "COCOLOG: A conditional observer and controller logic for finite machines," in *Proc. 29th IEEE Conf. Decis. Contr.*, Honolulu, HI, Dec. 1990, pp. 2845-2850.
- [3] P. E. Caines, T. Mackling, and Y. J. Wei, "Logic control via automatic theorem proving: COCOLOG fragments implemented in Blitzensturm 5.0," in *Proc. 1993 Amer. Contr. Conf.*, San Francisco, CA, June 1993, pp. 1209-1213.
- [4] Y. J. Wei and P. E. Caines, "On Markovian fragments of COCOLOG for logic control systems," in *Proc. 31st IEEE Conf. Decis. Contr.*, Tucson, AZ, Dec. 1992, 2967-2972.
- [5] S. Wang and P. E. Caines, "Automated reasoning with function evaluation for COCOLOG," INRIA-Sophia Antipolis, Research Rep. 1733, June 1992.
- [6] —, "Automated reasoning with function evaluation for COCOLOG with examples," in *Proc. 31st IEEE Conf. Decis. Contr.*, Tucson, AZ, Dec. 1992, pp. 3758-3763.
- [7] R. Goldblatt, *Logics of Time and Computation*. Stanford, CA: CSLI/Stanford, 1987.
- [8] J. H. Gallier, *Logic for Computer Science*. New York: Harper & Row, 1986.
- [9] T. Mackling, "The equality predicate and subsumption in resolution based automatic theorem proving," Dept. Elec. Eng., McGill Univ., Tech. Rep. 1993.
- [10] A. Haken, "The intractability of resolution," *Theoretical Computer Science*, vol. 39, pp. 297-308, 1985.

Robust Stability Criteria for Dynamical Systems Including Delayed Perturbations

Hansheng Wu and Koichi Mizukami

Abstract—In this note, we consider the problem of robust stability of uncertain time-delay dynamical systems. A new robust stability criteria for linear dynamical systems subject to delayed time-varying and nonlinear perturbations is derived. The results obtained in this note are less conservative than the ones reported so far in the literature. Some analytical methods are employed to investigate the bound on the perturbations so that the systems are stable. A numerical example is given to demonstrate the utilization of our results.

I. INTRODUCTION

It is well known that time delay is often encountered in various engineering systems, such as chemical processes, hydraulic, and rolling mill systems, and its existence is frequently a source of instability. Therefore, the stability problems of time-delay dynamical systems have received considerable attention over the decades (see, e.g., [1]–[8]). Because time-delay dynamical systems often include some perturbations, it is necessary to investigate the problems of robust stability of uncertain time-delay dynamical systems. In [9], for example, the problem of quantitative measures of robustness for linear dynamical systems including delayed nonlinear perturbations is investigated, and by using Razumikhin-type theorems, some bounds on the perturbations are obtained so that the systems remain stable. In [10], a delay-dependent stability condition for the linear uncertain time-delay dynamical systems is presented.

In this note, we consider the problem of robust stability criteria for linear dynamical systems subject to delayed time-varying and nonlinear perturbations. A new bound on the perturbations is presented so that the systems remain stable. Some analytical methods are employed to investigate this bound. Our results show that the robust stability criteria presented in this note is less conservative than the one given in [9]. In addition, since the robust stability condition developed in this note is independent of the delay, the results of this note may be applicable to dynamical systems with uncertain time-delay.

The rest of this note is organized as follows. In Section II, the problem to be tackled is precisely stated. In Section III, the main results of this note are proposed. As an application, an illustrative example is given in Section IV. The simulation of this example demonstrates the utilization of our results.

II. PROBLEM FORMULATION

We consider the linear dynamical systems including delayed perturbations described by the differential-difference equation of the form

$$\frac{dx}{dt} = Ax(t) + \Delta f(x(t-h(t)), t) \quad (1)$$

where $t \in R$ is the 'time,' $x(t) \in R^n$ is the current value of the state, $A \in R^{n \times n}$ is an asymptotically stable matrix, $\Delta f(\cdot, \cdot) : R^n \times R \rightarrow R^n$ is a time-varying nonlinear continuous function. The time-delay $h(t)$ is any nonnegative, bounded, and continuous function. That is, $h(t) \leq \bar{h}$, where \bar{h} is any constant.

Manuscript received February 10, 1993; revised April 29, 1993 and October 1993.

The authors are with the Division of Mathematical and Information Sciences, Faculty of Integrated Arts and Sciences, Hiroshima University, 1-7-1 Miyama, Higashi-Hiroshima 724, Japan.
E-Log Number 9407215.

Here, it is worth noting that we do not require for the magnitude of the time-varying delay $h(t)$ to be less than one, i.e., $h(t) < 1$. It is well known that such an assumption is often needed in papers dealing with stability problem of systems with time delay (see, e.g., [8]). Therefore, as will be seen in the next section, the results of this note may be applicable to the case where the time delay function is uncertain.

The initial condition for system (1) is given by

$$x(t) = \psi(t), \quad t \in [t_0 - \bar{h}, t_0] \quad (2)$$

where $\psi(\cdot)$ is a continuous function on $[t_0 - \bar{h}, t_0]$.

We assume that $\Delta f(\cdot, \cdot)$ represents the uncertainty and perturbations acting on system (1), and that there exists a nonnegative constant β such that for all $(x, t) \in R^n \times R$

$$\|\Delta f(x(t-h(t)), t)\| \leq \beta \|x(t-h(t))\| \quad (3)$$

where $\|\cdot\|$ denotes the Euclidean norm.

Now, the problem to be tackled in this note is to find some conditions on the perturbations so that system (1) is stable.

III. MAIN RESULTS

In this section, we derive the stability condition for system (1), which gives the bound on the perturbations. Before stating our main results we give the following definitions.

Definition 1. For any similarity transformation matrix M , we define

$$\mu = \|M\| \|M^{-1}\|. \quad (4)$$

Definition 2. For the asymptotically stable matrix A given in system (1), we define a constant α which satisfies

$$0 < \alpha < \min_i \{|\operatorname{Re} \lambda_i(A)|\}, \quad i = 1, \dots, n \quad (5)$$

where $\lambda_i(\cdot)$ denotes the i th eigenvalue of the square matrix (\cdot) , $\operatorname{Re} \lambda_i(\cdot)$ its real part.

Definition 3. For the symmetric positive definite matrix $P \in R^{n \times n}$, we let

$$\sigma = \frac{\lambda_{\max}(P)}{\lambda_{\min}(P)} \quad (6)$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximum eigenvalue and the minimum eigenvalue of matrix (\cdot) , respectively, and P satisfies the following Lyapunov-type equation

$$(\dot{A} + \alpha I)^T P + P(\dot{A} + \alpha I) = -Q \quad (7)$$

where I denotes the identity matrix, $Q \in R^{n \times n}$ is any symmetric positive definite matrix, and

$$\dot{A} = M^{-1} A M. \quad (8)$$

Remark 1: Since the matrix A is asymptotically stable, by Definition 1 the matrix $(\dot{A} + \alpha I)$ is also asymptotically stable. Therefore, in the light of the Lyapunov stability theory, Lyapunov-type (7) exists for any symmetric positive definite matrix Q a solution P which is also a symmetric positive definite matrix.

Then, for system (1), we have the following stability condition which results in a bound on the perturbations.

Theorem 1: If the following inequality is satisfied

$$\beta < \rho := \frac{\alpha}{\mu\sigma} \quad (9)$$

then system (1) is stable. More precisely, we have

$$\|z(t)\| \leq \frac{\sigma^{1/2}}{1-\bar{d}} \|D\| \exp\{-\bar{\gamma}(t-t_0)\} \quad (10)$$

where

$$\bar{d} := d(\bar{\gamma}) = e^{\bar{\gamma}\bar{h}} \frac{\mu\sigma\beta}{\alpha-\bar{\gamma}} < 1, \quad \bar{\gamma} \in [0, \alpha] \quad (11a)$$

$$\|D\| := \sup_{\tau \in [t_0-\bar{h}, t_0]} \|M^{-1}\psi(\tau)\| \quad (11b)$$

$$\|z(t)\| := \|M^{-1}\varphi(t)\| \quad (11c)$$

where M is any similarity transformation matrix.

Proof: We transform system (1) with the aid of a similarity transformation matrix M , and get

$$\frac{dz(t)}{dt} = Az(t) + M^{-1}\Delta f(Mz(t-h(t)), t) \quad (12)$$

with an initial condition

$$z(t) = M^{-1}\psi(t), \quad t \in [t_0-\bar{h}, t_0] \quad (13)$$

where $\psi = M^{-1}\varphi$.

Let $z(t)$ be the solution of system (12) with the initial condition (13) and define a positive definite function $V(z)$ as follows

$$V(z) = z^T(t)Pz(t). \quad (14)$$

In the light of Rayleigh principle [11], from (14) we have

$$\lambda_{\min}(P)\|z(t)\|^2 \leq V(z) \leq \lambda_{\max}(P)\|z(t)\|^2. \quad (15)$$

From (7), (12), and (14) we obtain

$$\begin{aligned} \frac{dV(z)}{dt} &= -2\alpha V(z) - z^T(t)Qz(t) \\ &\quad + 2z^T(t)PM^{-1}\Delta f(Mz(t-h(t)), t). \end{aligned} \quad (16)$$

Since both Q and P are positive definite, we also have

$$\frac{dV(z)}{dt} \leq -2\alpha V(z) + 2\mu\beta\lambda_{\max}(P)\|z(t)\|\|z(t-h(t))\|. \quad (17)$$

Therefore, from (17) we obtain the following inequality on $V(z)$

$$\begin{aligned} V(z) &\leq \lambda_{\max}(P)\|D\|^2 \exp\{-2\alpha(t-t_0)\} \\ &\quad + \int_{t_0}^t 2\mu\beta\lambda_{\max}(P) \exp\{-2\alpha(t-\tau)\} \\ &\quad \cdot \|z(\tau)\|\|z(\tau-h(\tau))\| d\tau. \end{aligned} \quad (18)$$

Here, we define an auxiliary function $S(t)$ as follows

$$S(t) = \begin{cases} \sqrt{\lambda_{\max}(P)}\|D\|, & t \in [t_0-\bar{h}, t_0] \\ \sqrt{\hat{S}(t)}, & t \geq t_0 \end{cases} \quad (19)$$

where

$$\begin{aligned} \hat{S}(t) &= \lambda_{\max}(P)\|D\|^2 \exp\{-2\alpha(t-t_0)\} \\ &\quad + \int_{t_0}^t 2\mu\beta\lambda_{\max}(P) \exp\{-2\alpha(t-\tau)\} \\ &\quad \cdot \|z(\tau)\|\|z(\tau-h(\tau))\| d\tau. \end{aligned}$$

Then, comparing (18) with (19) yields

$$S(t) \geq \sqrt{V(z)} \geq \sqrt{\lambda_{\min}(P)}\|z(t)\|. \quad (20)$$

That is

$$\frac{\|z(t)\|}{S(t)} \leq \frac{1}{\sqrt{\lambda_{\min}(P)}}. \quad (21)$$

Furthermore, from (19) we obtain

$$\begin{aligned} \frac{dS(t)}{dt} &= -\alpha S(t) + \mu\beta\lambda_{\max}(P) \frac{\|z(t)\|}{S(t)} \|z(t-h(t))\| \\ &\leq -\alpha S(t) + \frac{\mu\beta\lambda_{\max}(P)}{\sqrt{\lambda_{\min}(P)}} \|z(t-h(t))\|. \end{aligned} \quad (22)$$

Therefore, from (22) we also have

$$\begin{aligned} S(t) &\leq \sqrt{\lambda_{\max}(P)}\|D\| \exp\{-\alpha(t-t_0)\} \\ &\quad + \frac{\mu\beta\lambda_{\max}(P)}{\sqrt{\lambda_{\min}(P)}} \int_{t_0}^t \exp\{-\alpha(t-\tau)\} \\ &\quad \cdot \|z(\tau-h(\tau))\| d\tau. \end{aligned} \quad (23)$$

Comparing (20) with (23) yields

$$\begin{aligned} \|z(t)\| &\leq \sigma^{1/2}\|D\| \exp\{-\alpha(t-t_0)\} \\ &\quad + \mu\sigma\beta \int_{t_0}^t \exp\{-\alpha(t-\tau)\} \\ &\quad \cdot \|z(\tau-h(\tau))\| d\tau. \end{aligned} \quad (24)$$

Here, we first define the following continuous function

$$d(\gamma) := e^{\gamma\bar{h}} \frac{\mu\sigma\beta}{\alpha-\gamma}, \quad \gamma \in [0, \alpha].$$

It is obvious from the condition (9) that

$$d(0) = \frac{\mu\sigma\beta}{\alpha} < 1.$$

Then, according to the property of continuous function, there exists a constant $\bar{\gamma} > 0$ ($\bar{\gamma} < \alpha$) such that $d(\bar{\gamma}) < 1$. Here, for such a constant $\bar{\gamma} > 0$, we define

$$\bar{d} := d(\bar{\gamma}) < 1. \quad (25)$$

Continuing with (24), multiplying both sides of (24) by $\exp\{\bar{\gamma}(t-t_0)\}$ yields

$$\begin{aligned} \|z(t)\| \exp\{\bar{\gamma}(t-t_0)\} &\leq \sigma^{1/2}\|D\| \exp\{-\alpha(t-t_0) + \bar{\gamma}(t-t_0)\} \\ &\quad + \mu\sigma\beta \int_{t_0}^t \exp\{-\alpha(t-\tau) + \bar{\gamma}(t-t_0)\} \\ &\quad \cdot \|z(\tau-h(\tau))\| d\tau. \end{aligned} \quad (26)$$

With trivial manipulations, from (26) we obtain

$$\begin{aligned} \|z(t)\| \exp\{\bar{\gamma}(t-t_0)\} &\leq \sigma^{1/2}\|D\| + \mu\sigma\beta \int_{t_0}^t \exp\{\bar{\gamma}h(\tau)\} \\ &\quad \cdot \exp\{-(\alpha-\bar{\gamma})(t-\tau)\} \|z(\tau-h(\tau))\| \\ &\quad \cdot \exp\{\bar{\gamma}(\tau-h(\tau)-t_0)\} d\tau. \end{aligned} \quad (27)$$

Letting

$$Y_m(t) := \sup_{\rho \in [t_0-\bar{h}, t]} \|z(\rho)\| \exp\{\bar{\gamma}(\rho-t_0)\}$$

from (27) we have

$$\|z(t)\| \exp\{\bar{\gamma}(t-t_0)\} \leq \sigma^{1/2}\|D\| + e^{\bar{\gamma}\bar{h}} \frac{\mu\sigma\beta}{\alpha-\bar{\gamma}} Y_m(t). \quad (28)$$

In the light of the definition of $Y_m(t)$, from (25) and (28) we further obtain

$$\begin{aligned} Y_m(t) &\leq \sigma^{1/2}\|D\| + e^{\bar{\gamma}\bar{h}} \frac{\mu\sigma\beta}{\alpha-\bar{\gamma}} Y_m(t) \\ &= \sigma^{1/2}\|D\| + \bar{d} Y_m(t) \end{aligned}$$

i.e.,

$$Y_m(t) \leq \frac{\sigma^{1/2}}{1-\bar{d}} \|D\|. \quad (29)$$

TABLE I
THE COMPARISON OF THE RESULTS OBTAINED IN [9] AND THIS NOTE

	The Results of [9]	Our Results
Not Transformed	$\rho = 0.1458$	$\rho = 0.1835$
Transformed	$\rho = 0.1780$	$\rho = 0.2383$

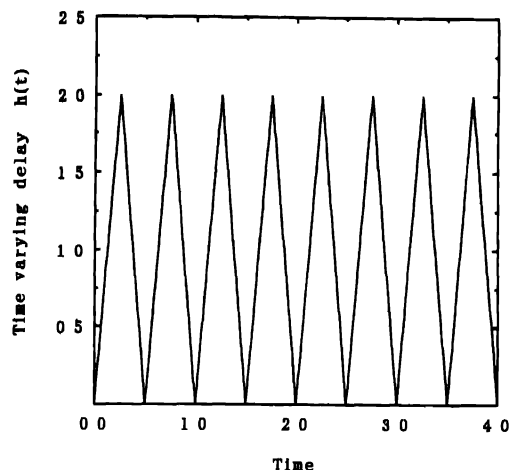


Fig 1 Time delay history

From (29) we can obtain that for $t \geq t_0$

$$\|x(t)\| \leq \|x(t_0)\| \exp\{-\gamma(t-t_0)\} \leq \frac{\sigma^{1/\alpha}}{1-\bar{\alpha}} \|D\| \exp\{-\gamma(t-t_0)\} \quad (30)$$

It is obvious from (30) that system (12) is exponentially stable. This implies [noting $\bar{x} = M^{-1}x$] that system (1) is also exponentially stable. \square

Remark 2 The robust stability condition (9) gives a bound on the perturbations. It is worth noting that (9) is independent of delay. Therefore the results obtained here is applicable to the case where the delay function $h(t)$ is uncertain.

Remark 3 The robust stability condition (9) does not include $\lambda(C)/\lambda_{\min}(P)$ which is included in the bound given in [9]. Then our results may be less conservative than the ones given in [9]. This will further be illustrated in the following numerical example.

IV. NUMERICAL EXAMPLE

We recall from [9] the following matrix

$$A = \begin{bmatrix} -3 & -2 \\ 1 & 0 \end{bmatrix} \quad (31)$$

In [9] selecting the similarity transformation matrix M and Q as identity matrix I yields for the perturbations the bound $\rho = 0.158$. Further improvement is obtained by applying the following transformation suggested in [12] for the perturbations

$$M = \begin{bmatrix} 0.99964 & -0.26217 \\ 0.02660 & 0.95937 \end{bmatrix} \quad (32)$$

yields a robust bound $\rho = 0.178$.

Our results this bound can further be improved. Here, if we select $\alpha = 0.5$, $M = I$, and

$$Q = \begin{bmatrix} 15.000 & 3.000 \\ 3.000 & 2.647 \end{bmatrix} \quad (33)$$

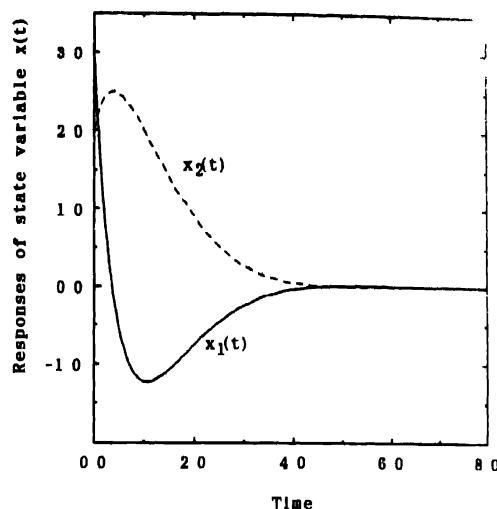


Fig 2 Simulation of example for $\gamma = 0.22$

then we can have $\rho = 0.1835$. Furthermore, we also employ the transformation (32) and let

$$Q = \begin{bmatrix} 8.000 & 1.000 \\ 1.000 & 0.126 \end{bmatrix} \quad (34)$$

then we can obtain a bound $\rho = 0.23893$ which improves the bound reported in [9] for the perturbations (by 34%).

The comparison of our results with the ones obtained in [9] is shown in Table I. It is obvious from Table I that the results obtained in [9] are more conservative than our ones. To further confirm this, the simulation of this example for $\gamma = 0.22$ ($< \rho = 0.2383$) is depicted in Fig 2. Here we have employed for the simulation the following

$$\Delta f(\bar{x}) = \begin{bmatrix} 0.20 \sin[r_2(t-h(t))] \\ 0.22 \sin[r_1(t-h(t))] \end{bmatrix}$$

$$r_1(t) = [3 \cos(t) \quad 2 \cos(t)]^T$$

and the time varying delay $h(t)$ is given in Fig 1. Note that $h(t)$ is not defined at $t = 0.25n$, $n = 1, 2, \dots$

It is shown from Fig 2 that the system is indeed exponentially stable.

REFERENCES

- [1] W. H. Kwon and A. F. Pearson, Feedback stabilization of linear systems with delay control, *IEEE Trans Automat Contr*, vol AC 25, pp 266-269, 1980.
- [2] N. K. Patel, P. C. Das, and S. S. Prabhu, Optimal control of systems described by delay differential equations, *Int J Contr*, vol 36, pp 303-311, 1982.
- [3] R. I. Alford and E. B. Lee, Sampled data hereditary systems: Linear quadratic theory, *IEEE Trans Automat Contr*, vol AC 31, pp 60-65, 1986.
- [4] K. Watanabe, M. Ito, M. Kaneko, and T. Ouchi, Finite spectrum assignment problem for systems with delay in state variable, *IEEE Trans Automat Contr*, vol AC 28, pp 506-508, 1983.
- [5] A. Thowsen, Stabilization of a class of linear time delay systems, *Int J Systems Sci*, vol 12, pp 1485-1492, 1981.
- [6] A. Felachi and A. Thowsen, Memoryless stabilization of linear delay differential systems, *IEEE Trans Automat Contr*, vol AC 26, pp 586-587, 1981.
- [7] T. More, E. Noldus, and M. Kuwahara, A way to stabilize linear systems with delayed state, *Automatica*, vol 19, pp 571-573, 1983.
- [8] I. Ikeda and T. Ashida, Stabilization of linear systems with time-varying delay, *IEEE Trans Automat Contr*, vol AC 24, pp 369-370, 1979.
- [9] F. Cheres, Z. J. Palmor, and S. Gutman, Quantitative measures of robustness for systems including delayed perturbations, *IEEE Trans Automat Contr*, vol AC 34, pp 1203-1205, 1989.

- [10] T. J. Su and C. G. Huang, "Robust stability of delay dependence for linear uncertain systems," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1656-1659, 1992.
- [11] J. N. Franklin, *Matrix Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1968.
- [12] R. K. Yedavalli and Z. Liang, "Reduced conservatism in stability robustness bounds by state transformation," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 863-866, 1986.

An Efficient Method for Unconstrained Optimization Problems of Nonlinear Large Mesh-Interconnected Systems

Shin-Yeu Lin and Ch'i-Hsin Lin

Abstract—We present a new efficient method for solving unconstrained optimization problems for nonlinear large mesh-interconnected systems. This method combines an approximate scaled gradient method with a block Gauss-Seidel with line search method which is used to obtain an approximate solution of the unconstrained quadratic programming subproblem. We prove that our method is globally convergent and demonstrate by several numerical examples its superior efficiency compared to a sparse matrix technique based method. In an example of a system of more than 200 variables, we observe that our method is 3.45 times faster than the sparse matrix technique based Newton-like method and about 50 times faster than the Newton-like method without the sparse matrix technique.

I. INTRODUCTION

In this paper we consider the following unconstrained optimization problem for a nonlinear large mesh-interconnected system

$$\min_{x \in \mathbb{R}^n} J(x) \quad (1)$$

where the objective function $J: \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable, bounded from below and satisfies the Lipschitz condition that there exists a constant $K > 0$ such that $\|\nabla J(x_1) - \nabla J(x_2)\|_2 \leq K\|x_1 - x_2\|_2, \forall x_1, x_2 \in \mathbb{R}^n$. A general descent algorithm, which is called a scaled gradient method in [1], for solving problem (1) uses the following iterations

$$x^{k+1} = x^k + \gamma^k s^k \quad (2)$$

where k denotes the iteration index, γ^k is a step-size, and s^k is the solution of

$$\min_{s \in \mathbb{R}^n} \frac{1}{2} s^T C(x^k) s + \nabla J(x^k)^T s \quad (3)$$

in which $C(x^k)$ is a positive definite matrix.

For a large mesh-interconnected system, if $C(x^k)$ is selected so that (2) is a Newton or Newton-like method, $C(x^k)$ may be a sparse matrix. The solution of (3), which is the solution of the linear system

$$C(x^k) s = -\nabla J(x^k) \quad (4)$$

can be obtained using the sparse matrix technique, which is a very powerful means for solving linear equations in a circuit system [2] or

an electric power system [3]. This technique effectively reduces the amount of computer memory needed and improves the computation time dramatically because it stores only nonzero elements and ignores operations involving zeros in the solution process [2]. Thus, a sparse matrix technique based Newton or Newton-like method is much more efficient than a method with quadratic convergence rate.

In this paper, we will present a new method for solving (1) for a large mesh-interconnected system to compete with a sparse matrix technique based method. Our method combines an approximate scaled gradient method with a block Gauss-Seidel with line search (BGSLS) method. Although each of the above methods is well known, combining them into one method is a new approach. The basic idea behind this method is to preserve the advantages and discard the disadvantages of the block Gauss-Seidel method. We have observed in many numerical experiments [4] that the block Gauss-Seidel method approaches its convergence point very fast in the first few iterations and then slows down around that point. This fact indicates that the block Gauss-Seidel method is better suited for use as a descent direction generator rather than as an optimization algorithm by itself. To improve the quality of the descent direction further, we use an exact line search at the end of each cycle of the block Gauss-Seidel method and thus form a BGSLS method. Executing the BGSLS method for a finite number of iterations with appropriate stopping criteria will generate the descent direction needed for the approximate scaled gradient method. We will show in this paper that our method is a globally convergent descent algorithm. It is difficult to give an analytical convergence rate for our method because of the nature of the method. However, since the major computation required by our method lies in the execution of the BGSLS method, our method will be efficient if the BGSLS method can generate an effective descent direction in several iterations. Intuitively, our method should be efficient and effective, for two reasons: i) only small minimization problems are involved in each iteration of the BGSLS method and ii) the exact line search used at the end of each iteration of the BGSLS method greatly improves the quality of the descent direction.

II. SOLUTION METHOD

A. The Approximate Scaled Gradient Method

The approximate scaled gradient method [1] is

$$x^{k+1} = x^k + \gamma^k s^k \quad (5)$$

where s^k is an approximate solution of (3) and γ^k is a step-size. To ensure global convergence, we use Armijo-type rule to determine the step-size by

$$\gamma^k = \beta^{m_k} \lambda \quad (6)$$

where $0 < \beta < 1$, $\lambda > 0$, and m_k is the smallest nonnegative integer such that makes the following inequality hold for some positive constant K_2

$$J(x^k + \beta^{m_k} \lambda s^k) - J(x^k) \leq -\frac{K_2}{2} \beta^{m_k} \lambda \|s^k\|_2^2. \quad (7)$$

Remark 2.1: a) Condition (7) is to ensure a sufficient decrement of the objective function obtained in each iteration of (5); the satisfaction of this condition serves as a terminating criteria for our Armijo-type step-size rule (6). b) As will be shown in Theorem 2.1, the matrix $\frac{1}{2} C(x^k) - K_2 I$ being nonnegative definite is a sufficient condition for (5) to converge. This sufficient condition provides a value for λ .

Manuscript received October 19, 1993; revised March 8, 1994. This research work is supported by National Science Council in R.O.C. under Grants NSC-82-0404-E-009-166 and NSC-79-0404-E-009-48.

The authors are with the Department of Control Engineering, National Chiao Tung University, Hsinchu, Taiwan.

IEEE Log Number 9407217.

Lemma 2.1: Suppose (3) is solved by a descent iterative algorithm starting from $s = 0$ for any arbitrary number of iterations, and let \hat{s}^k denote the final value of s . Then $\nabla J(x^k)^T \hat{s}^k < 0$.

This lemma can easily be verified by the fact that $\frac{1}{2}(\hat{s}^k)^T C(x^k) \hat{s}^k - \nabla J(x^k)^T \hat{s}^k < 0$. Thus, our approximate scaled gradient method (5) will be an efficient descent algorithm for solving (1) if the descent algorithm we employ to obtain \hat{s}^k is efficient.

B Block Gauss-Seidel with Line Search (BGSLS) Method

1) **One Block Gauss-Seidel Cycle:** Let us partition s into p subvectors such that $s = [s_1 s_2 \cdots s_p]^T$. Then one block Gauss-Seidel cycle is to perform

$$\min_{s_i} \frac{1}{2} s^T C(x^k) s + \nabla J(x^k)^T s \quad (8)$$

from $i = 1$ to p . In (8), the subvector s_i is taken as the vector of minimizing variables while the variables in the subvectors $s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_p$ are held fixed at their current values. Note that compared to (3), (8) is a small unconstrained minimization problem for every i .

Remark 2.2. On the partition of the s -vector, there are two extremes corresponding to $p = 1$ and $p = n$. The case of $p = 1$ is not the interest of this paper. For the case of $p = n$, the descent direction generated by (8) has very poor quality. Thus, a good partition should take the following two factors into account: a) computational burden of solving (8) and b) the quality of descent direction generated. In fact, the above two factors have conflicting interests; at this stage, we have not yet achieved an optimal way to partition the s -vector. Nonetheless, for a network-structure like system, it would be beneficial if each partitioned subnetwork is mesh-interconnected, and the sizes of all subnetworks do not differ much.

2) **Optimal Step-Size** Let $s^{(i)}$ denote the value of s after solving (8) for s_i . Then $s^{(0)}$ represents the initial value and $s^{(p)}$ the final value of s for one block Gauss-Seidel cycle. Suppose $s^{(0)}$ is not the optimal solution of (3). Then the following inequality holds

$$\begin{aligned} \frac{1}{2} (s^{(0)})^T C(x^k) s^{(0)} + \nabla J(x^k)^T s^{(0)} \\ > \frac{1}{2} (s^{(p)})^T C(x^k) s^{(p)} + \nabla J(x^k)^T s^{(p)}. \end{aligned} \quad (9)$$

Furthermore, because $C(x^k)$ is positive definite, $\frac{1}{2} s^T C(x^k) s + \nabla J(x^k)^T s$ is a convex function in s . Based on this fact, we may verify that

$$\begin{aligned} \frac{1}{2} (s^{(p)})^T C(x^k) s^{(p)} + \nabla J(x^k)^T s^{(p)} \\ \geq \frac{1}{2} (s^{(0)})^T C(x^k) s^{(0)} + \nabla J(x^k)^T s^{(0)} \\ + [C(x^k) s^{(0)} + \nabla J(x^k)]^T (s^{(p)} - s^{(0)}). \end{aligned} \quad (10)$$

Then, combining (9) and (10), we obtain the following lemma.

Lemma 2.2: Let $s^{(0)}$ and $s^{(p)}$ denote the initial and final values of s -variables, respectively, for one block Gauss-Seidel cycle of the BGSLS method. Suppose $s^{(0)}$ is not an optimal solution of (3). Then

$$[C(x^k) s^{(0)} + \nabla J(x^k)]^T (s^{(p)} - s^{(0)}) < 0. \quad (11)$$

The above inequality implies that $s^{(p)} - s^{(0)}$ is a descent direction of $\frac{1}{2} s^T C(x^k) s + \nabla J(x^k)^T s$ at $s = s^{(0)}$.

Therefore, we can determine the exact optimal step-size $\hat{\alpha}$ to update variable s by

$$s'' = s^{(0)} + \hat{\alpha} (s^{(p)} - s^{(0)}) \quad (12)$$

where s'' denotes the updated s , and the optimal step-size

$$\hat{\alpha} = - \frac{[C(x^k) s^{(0)} + \nabla J(x^k)]^T (s^{(p)} - s^{(0)})}{[s^{(p)} - s^{(0)}]^T C(x^k) [s^{(p)} - s^{(0)}]} \quad (13)$$

is obtained by solving the following one-dimensional minimization problem

$$\begin{aligned} \min_{\alpha \geq 0} \{ & \frac{1}{2} [s^{(0)} + \alpha (s^{(p)} - s^{(0)})]^T C(x^k) [s^{(0)} + \alpha (s^{(p)} - s^{(0)})] \\ & + \nabla J(x^k)^T [s^{(0)} + \alpha (s^{(p)} - s^{(0)})] \}. \end{aligned}$$

3) **One Iteration of the BGSLS Method.** The following three operations form one iteration of the BGSLS method: i) execute one block Gauss-Seidel cycle; ii) determine α ; and iii) update s'' . s'' will be the initial value $s^{(0)}$ for the next iteration of the BGSLS method.

4) **Convergence of the BGSLS Method:** This iterative BGSLS method will converge to the solution of (3), as described in the following lemma. The proof of the lemma is given in the Appendix.

Lemma 2.3: Assuming that there exists a constant $K_2 > 0$ such that $\frac{1}{2} C(x^k) - K_2 I$ is nonnegative definite for all x^k , then a) the BGSLS method is a descent method, and b) any limit point of the sequence generated by the BGSLS method is a solution of (3).

5) **Stopping Criteria of the BGSLS Method:** As pointed out in Section I, the BGSLS method approaches a solution point of (3) very quickly in the first few iterations; however, it then slows down around that point. In fact, it was this characteristic of the BGSLS method that gave us the idea of combining it with the approximate scaled gradient method. Therefore, to obtain the \hat{s}^k needed in (5), we do not need to execute the BGSLS method until it converges. In fact, we can stop the BGSLS method after a finite number of iterations. Consequently, one of the stopping criteria of this method is if the improvement of the objective function satisfies

$$\frac{Q(t) - Q(t+1)}{Q(t)} \times 100\% < \xi\% \quad (15)$$

where $Q(t) = \frac{1}{2} s(t)^T C(x^k) s(t) + \nabla J(x^k)^T s(t)$, t denotes the iteration index of the BGSLS method, and $\xi\%$ is a preselected percentage. To ensure that (5) converges, another stopping criteria

$$\|s(t+1)\|_2 < K_1 \|\nabla J(x^k)\|_2 \quad (16)$$

for some $K_1 > 0$, for all $t > t'$, where t' is a finite positive integer, should also be satisfied. Then, the $s(t+1)$ which satisfies stopping criteria (15) and (16) will be set as \hat{s}^k . An explanation of the need for (16) will be given in the proof of Theorem 2.1. The existence of K_1 and t' is ensured by the following corollary, the proof of which is also given in the Appendix.

Corollary 2.1: Let $\{s(t)\}$ denote a sequence generated by the BGSLS method. Under the assumption of Lemma 2.3, there exists a t' such that $\|s(t)\|_2 > K_1 \|\nabla J(x^k)\|_2$ for some $K_1 > 0$ and for all $t > t'$.

C The Algorithm for the Solution Method and its Convergence

Combining Lemma 2.1 and Lemma 2.3 then proves that our method is a descent method.

Lemma 2.4: The combination of the approximate scaled gradient method (5) and the BGSLS method, supplemented by the stopping criteria (15) and (16), is a descent method for solving (1).

1) **The Algorithm:** We are now ready to state our algorithm as follows: *Given data:* i) $s = [s_1, s_2, \dots, s_p]^T$, ii) the values of $\lambda (> 0)$ and $\beta (0 < \beta < 1)$ in Armijo-type rule, iii) the value of $\xi\% (\geq 0\%)$ in (15), and iv) a positive constant K_2 which meets the assumption in Lemma 2.3.

Step 0: Set the values $x^{(0)}$ and set $k = 0$.

Step 1: Set $t = 1$ and set $s(t) = 0$.

Step 2: Set $i = 1$ and $s^{(0)}(t) = s(t)$.

Step 3: Solve (8) to obtain $s^{(i)}$. If $i = p$, go to Step 5; otherwise, set $i = i + 1$ and repeat this step.

Step 4: Compute $s^{(p)}(t) - s^{(0)}(t)$ and determine $\hat{\alpha}(t)$ by (13).

Step 5: Update $s(t+1) = s^{(0)}(t) + \hat{\alpha}(t)[s^{(p)}(t) - s^{(0)}(t)]$.

Step 6: If (15) and (16) are satisfied, set $\hat{s}^k = s(t+1)$ and go to Step 7; otherwise, set $t = t+1$ and return to Step 2.

Step 7: Set $m = 0$.

Step 8: If $J(x^k + \gamma^m \lambda s^k) - J(x^k) \leq -(K_2/2) \gamma^m \lambda \|\hat{s}^k\|_2^2$, set $\gamma^k = \gamma^m \lambda$; otherwise, set $m = m+1$, and repeat this step.

Step 9: Update $x^{k+1} = x^k + \gamma^k \hat{s}^k$, set $k = k+1$, and return to Step 1.

2) Convergence of the Algorithm: The following theorem ensures the convergence of our algorithm. A proof of the theorem is given in the Appendix.

Theorem 2.1: Let $\{x^k\}$ denote the sequence generated by our algorithm. Under the assumption of Lemma 2.3, $\lim_{k \rightarrow \infty} \nabla J(x^k) = 0$.

Remark 2.3: The values of λ , β , and $\xi\%$ do not affect the convergence but will influence the efficiency of our method. Thus, these values can be determined empirically for individual systems.

3) Stopping Criteria of the Algorithm: From Theorem 2.1, we see that as $k \rightarrow \infty$, $\hat{s}^k \rightarrow 0$; however, for practical considerations, our algorithm will stop when $\|\hat{s}^k\|_\infty < \epsilon$ for a reasonable accuracy.

III. EXAMPLES

Most electric power systems are nonlinear large mesh-interconnected systems. The weighted least squares problem in power system state estimation is a typical unconstrained optimization problem and thereby is an adequate example for demonstrating the computational efficiency of our method.

A. The Weighted Least Squares Problem in Power System State Estimation

The power system state estimation problem [6] can be briefly described as follows: For an l -bus¹ power system, the voltage magnitudes and phase angles of all buses constitute the states of the system. Because the bus phase angle cannot be measured, the states of the system have to be estimated based on available measurements, which are functions of the states, such as power transmission line flow, bus voltage magnitude, bus power injection, and transformer tap. All such measurements can be expressed in a general form as $z = h(x) + \eta$, where x is a $2l$ -dimensional vector of state variables, z is an m -dimensional vector of measurements, h denotes m nonlinear measurement functions which are twice continuously differentiable, and η represents an m -dimensional Gaussian random vector of measurement errors with an $m \times m$ diagonal covariance matrix R . A common formulation [5] of this state estimation problem is to solve the following weighted least squares problem²

$$\min_x J(x) (= \frac{1}{2} [z - h(x)]^T R^{-1} [z - h(x)]). \quad (17)$$

Since $h(x)$ is twice continuously differentiable, $J(x)$ in (17) is twice continuously differentiable and thereby satisfies Lipschitz condition. Taking $n = 2l$, the $J(x)$ meets the assumptions given in Section I.

Note: For an authentic power system state estimation [6], bad data processing has to be associated with the solution of (17); if bad measurements are present, they should be eliminated, and the states re-estimated. However, for the purposes of this paper, we will focus on solving (17) only.

¹ The buses of the power systems considered here are similar to nodes in an electrical network.

² The phase angle of the slack bus is considered to be a known value.

B. Conventional Approach

In general, the Newton method may fail to solve (17) because the Hessian matrix of (17) may not be positive definite. Thus, the conventional approach in the power system literature [6] uses a Newton-like method, which is (2) with $C(x^k) = 2H(x^k)^T R^{-1} H(x^k)$ in (3) where $H(x^k) = \nabla h(x^k)$. However, $H(x^k)^T R^{-1} H(x^k)$ may be singular or ill-conditioned, since it is only positive semidefinite. To cope with this difficulty, we can modify the above $C(x^k)$ to ensure the positive definiteness by setting $C(x^k) = 2H(x^k)^T R^{-1} H(x^k) + \delta I$, where δ is a small positive real constant and I is the identity matrix. We see that under this modification, the only assumption needed in Theorem 2.1, $\frac{1}{2} C(x^k) - K_2 I$ is nonnegative definite, is satisfied by taking $K_2 = (\delta/2)$. Furthermore, the numerical stability of our algorithm is guaranteed if δ is not too small. Consequently, (4) becomes

$$[2H(x^k)^T R^{-1} H(x^k) + \delta I]s = -\nabla J(x^k). \quad (18)$$

The matrix $2H(x^k)^T R^{-1} H(x^k) + \delta I$ and the matrix $H(x^k)^T R^{-1} H(x^k)$ are sparse, and (18) can be solved using the sparse matrix technique.

Remark 3.1: a) Because $H(x^k)^T R^{-1} H(x^k)$ may be ill conditioned, methods that use an orthogonal transformation technique to solve $2H(x^k)^T R^{-1} H(x^k)s = -\nabla J(x^k)$ has been developed [7]. However, these methods are computationally very inefficient. b) In [6], the step-size γ^k is set to be 1 for all k . In fact, with a constant step-size such as this there is no guarantee that (2) will converge. c) There are other methods in the power system literature that can obtain an approximate solution for state estimation using a decoupling approach [8]. However, these methods are beyond the scope of this paper, since we are concerned only with methods that solve (1) exactly.

C. Application of the Proposed Method

Let us partition the power system network into $p (> 1)$ subnetworks such that each subnetwork is a connected graph in the topological sense. Let B_i denote the set of buses of the i th subnetwork; then s_{B_i} denotes the subvector of s corresponding to states of the i th subnetwork. Replacing s_i , $C(x^k)$, and $\nabla J(x^k)$ in the proposed algorithm by s_{B_i} , $2H(x^k)^T R^{-1} H(x^k) + \delta I$, and $H(x^k)^T R^{-1} (z - h(x^k))$, respectively, we can apply the proposed algorithm directly to (17). Analogous to (8), we see that $s_{B_1}, \dots, s_{B_{i-1}}, s_{B_{i+1}}, \dots, s_{B_p}$ are held fixed in the i th minimization problem

$$\min_{s_{B_i}} \frac{1}{2} s^T C(x^k) s + \nabla J(x^k)^T s \quad (19)$$

whose solution can be obtained by solving a small set of linear equations involving only the i th subnetwork and the boundary buses outside the i th subnetwork [9].

D. Test Results

We applied our method, the sparse matrix technique based Newton-like method, and the Newton-like method without sparse matrix technique to the weighted least squares problem in state estimation of the IEEE 30-bus power system and the IEEE 118-bus power system. For the cases under consideration, the IEEE 30-bus system was partitioned into three subnetworks, as shown in Fig. 1, in which each subnetwork is indicated by closed dashed contours. The IEEE 118-bus system is partitioned into eight subnetworks. Because of the limitation on the length of this paper, we can not include here a figure of the IEEE 118-bus system, however, this figure can be found in [10]. Nonetheless, we list the bus number of the buses of each subnetwork in the following. Subnetwork 1 contains buses 1 to 12.

TABLE I
COMPARISON OF OUR METHOD WITH NEWTON-LIKE METHOD WITH AND WITHOUT SPARSE MATRIX TECHNIQUE

System	Final objective value			Average CPU time (second:)			Speedup ratio	
	NLWTS ^(Δ) method	NLWS ^(□) method	Our method	NLWTS ^(Δ) method (I)	NLWS ^(□) method (II)	Our method (III)	$\frac{1}{T}$	$\frac{1}{T/I}$
IEEE-30 bus	15.88	15.88	15.88	0.99	0.46	0.27	2.15	3.67
IEEE-118 bus	43.05	43.05	43.05	76.53	5.49	1.59	13.94	48.13

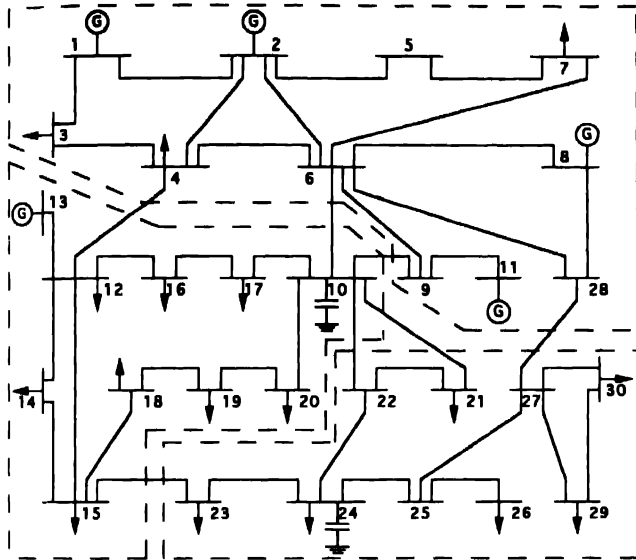


Fig. 1. The IEEE 30-bus system and three partitioned subnetworks.

and bus 117. Subnetwork 2 contains buses 15 to 19, buses 27 to 32, and buses 113 and 114. Subnetwork 3 contains buses 20 to 26, buses 70 to 76, and bus 118. Subnetwork 4 contains buses 33 to 47. Subnetwork 5 contains buses 50 to 61, and buses 63 and 64. Subnetwork 6 contains buses 48, 49, and 62, buses 65 to 69, buses 77 to 81, and buses 97, 98, and 116. Subnetwork 7 contains buses 82 to 96. Subnetwork 8 contains buses 99 to 112. Note that each subnetwork contains about 10 to 20 buses in mesh-interconnected structure though the bus numbers are not consecutive.

In both cases, we took the real-power flow and the reactive-power flow of all transmission lines of each individual system as the measurement data and assumed that some of these are bad data. We set the parameters in our algorithm as follows: $\delta = 0.01$, $\beta = 0.9$ for both systems, $\xi = 50\%$ for the IEEE 30-bus system, and $\xi = 10\%$ for the IEEE 118-bus system, $\lambda = 1.05, 1.10, 1.15$ for the IEEE 30-bus system, and $\lambda = 1.25, 1.30, 1.35$ for the IEEE 118-bus system. Note that each value of λ corresponds to one computer run. We used various values of λ to test Armijo-type rule and average the CPU time. The accuracy ϵ in the stopping criterion of our algorithm, which is described in Subsection II-C-3), was set to be 10^{-2} for both cases. We used a Sun 4/60 workstation to test our algorithm. The simulation results for the case of the IEEE 30-bus system and the use of the IEEE 118-bus system are shown in Table I. The average CPU time of our algorithm is the average CPU time of the outer runs with various values of λ reported on the Sun 4/60 workstation. We also solved the same cases for both systems using the sparse-matrix technique based Newton-like method, with the same setup of Armijo-type rule and same initial guess as our method on the same Sun 4/60 workstation. We used $\|s^k\|_\infty < \epsilon = 10^{-2}$ as the stopping criteria for the sparse matrix technique based Newton-

like method, the test results for which are also shown in Table I. We also applied the Newton-like method without using the sparse matrix technique to the same cases on the same workstation with the same setup of Armijo-type rule, same initial guess, and same stopping criteria as the sparse matrix technique based Newton-like method. The resulting average CPU times are also reported in Table I for the cases of both systems. We see that our algorithm achieved the same final objective value as the Newton-like methods. As expected, the sparse matrix technique based Newton-like method was much faster than the Newton-like method without the sparse matrix technique in both cases, especially in the case of the IEEE 118-bus system, which has more sparsity. Compared to the sparse matrix technique based Newton-like method, our method also performs better with the larger system. From the speedup ratio shown in Table I, we see that in the case of the IEEE 118-bus system, our method is 3.45 times faster than the sparse matrix technique based Newton-like method and 48.13 times faster than the Newton-like method without the sparse matrix technique. This demonstrates the dramatic increase in efficiency provided by our method. In order to better appreciate the merits of our algorithm, in Fig. 2 we describe the details of the progression of our algorithm and the Newton-like method with the sparse matrix technique in solving the case of the IEEE 118-bus system when $\lambda = 1.35$. The result of our algorithm is shown by the curve marked with circles \circ and associated with the Arabic numeral iteration index. Each circle indicates the CPU time accumulated (horizontal axis) at that iteration versus the corresponding value of the objective function (vertical axis) in the progression. The curve marked with asterisks $*$ and associated with the Roman numeral iteration index corresponds to the Newton-like method with the sparse matrix technique. Our algorithm takes 12 iterations to meet the stopping criteria $\|s^k\|_\infty < 10^{-2}$, while the sparse matrix technique based Newton-like method takes only 11 iterations to meet the stopping criteria $\|s^k\|_\infty < 10^{-2}$. However, we see that the amount of CPU time taken up by each iteration of our algorithm is far less than that for each iteration of the Newton-like method with the sparse matrix technique. This shows the effectiveness and efficiency of using the BGSLs method to generate the descent direction. Furthermore, when our algorithm has nearly reached the optimal objective value, the Newton-like method with the sparse matrix technique has just finished its first iteration and the corresponding objective value is still quite far away from the optimal value. The efficiency of our algorithm is obvious.

IV. CONCLUSION

We have developed a globally convergent algorithm for unconstrained optimization problems of nonlinear large mesh-interconnected systems. Although, due to the nature of our method, an analytical convergence rate is not available, the dramatic increase in efficiency provided by our method can be observed both from the method itself and from the simulation results. It is worth noting that the BGSLs method can be processed by parallel processors if the order of the partitioned subnetworks is suitably arranged [11]. Thus, the speed of our method may be further increased by using parallel processors.

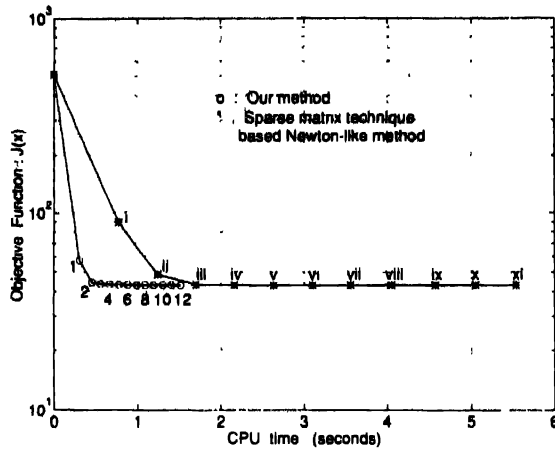


Fig. 2. Details of the progression of our method and the sparse matrix technique based Newton-like method in solving the weighted least squares problem for the IEEE 118-bus system.

APPENDIX

Proof of Lemma 2.3: a) Let $T = \{s \in \mathbb{R}^{2n} | C(x^k)s + \nabla J(x^k) = 0\}$ denote the solution set of (3). Then if $s \notin T$, the descent property of the BGSLs method follows directly from Lemma 2.2 and (12)–(14). According to the Global Convergence Theorem [12, pp. 187–188], b) can be proven if we show that i) the BGSLs method is a descent method, ii) the sequence generated by the BGSLs method lies in a compact set, and iii) the mapping of any iteration of the BGSLs method is closed. Clearly, i), which is a), has been shown. In the following, we will prove ii) and iii).

Because $C(x^k)$ is positive definite, the unique minimum solution of (3) is

$$s^* = -C(x^k)^{-1} \nabla J(x^k). \quad (A1)$$

Let the set $S \equiv \{s \in \mathbb{R}^{2n} | -\frac{1}{2} \nabla J(x^k)^T C(x^k)^{-1} \nabla J(x^k) \leq \frac{1}{2} s^T C(x^k) s + \nabla J(x^k)^T s \leq 0\}$. Then $S \neq \emptyset$, since $0 \in S$. We claim that S is compact. Clearly, S is closed. Thus, it is enough to show that S is bounded. Suppose not, $\exists s \in S$ \exists

$$\|s\|_2 > \max \left(\frac{\|\nabla J(x^k)\|_2 + 1}{K_2}, \left| \frac{1}{2} \nabla J(x^k)^T C(x^k)^{-1} \nabla J(x^k) \right| + 1 \right). \quad (A2)$$

For this s and by the assumption that $\frac{1}{2} C(x^k) - K_2 I > 0$, we have

$$\begin{aligned} & \left| \frac{1}{2} s^T C(x^k) s + \nabla J(x^k)^T s \right| \\ & \geq \left| \frac{1}{2} s^T C(x^k) s \right| - |\nabla J(x^k)^T s| \\ & \geq K_2 \|s\|_2^2 - \|\nabla J(x^k)\|_2 \|s\|_2 \\ & = (K_2 \|s\|_2 - \|\nabla J(x^k)\|_2) \|s\|_2. \end{aligned} \quad (A3)$$

From (A2)

$$\begin{aligned} & (K_2 \|s\|_2 - \|\nabla J(x^k)\|_2) \|s\|_2 > \|s\|_2 \\ & > \left| \frac{1}{2} \nabla J(x^k)^T C(x^k)^{-1} \nabla J(x^k) \right|. \end{aligned} \quad (A4)$$

Thus, from (A3) and (A4), we obtain that $\left| \frac{1}{2} s^T C(x^k) s + \nabla J(x^k)^T s \right| > \left| \frac{1}{2} \nabla J(x^k)^T C(x^k)^{-1} \nabla J(x^k) \right|$. This inequality contradicts $s \in S$. Hence S must be bounded. From (a), the BGSLs method is a descent method. Thus, every point in the sequence $\{s(t)\}$ generated by the BGSLs method starting from $s = 0$ should satisfy

$$\frac{1}{2} s(t)^T C(x^k) s(t) + \nabla J(x^k)^T s(t) \leq 0, \quad \forall t \quad (A5)$$

moreover, by the descent property

$$\begin{aligned} & \frac{1}{2} s^T C(x^k) s + \nabla J(x^k)^T s \\ & \leq \frac{1}{2} s(t)^T C(x^k) s(t) + \nabla J(x^k)^T s(t), \quad \forall t. \end{aligned} \quad (A6)$$

From (A1) the left-hand side of (A6) equals $-\frac{1}{2} \nabla J(x^k)^T C(x^k)^{-1} \nabla J(x^k)$, thus from (A5) and (A6), the sequence $\{s(t)\}$, $\forall t$, must lie inside the compact set S . This proves ii). Next, we will prove iii). Based on the fact that the composition of closed mappings is a closed mapping, and the exact line search method is also a closed mapping [12, p. 210], it is enough to show iii) if we can show that the mapping of Step 3 of our algorithm is closed. This is true because (8) is a bounded, unconstrained quadratic minimization problem with positive definite $C(x^k)$. This completes the proof. \square

Proof of Corollary 2.1: Since by Lemma 2.3, the sequence $\{s(t)\}$ converges to a point s^* satisfying $C(x^k)s^* + \nabla J(x^k) = 0$, then $\|C(x^k)\|_2 \|s^*\|_2 \geq \|\nabla J(x^k)\|_2$. Let $K_0 = (1/\sup_k \|C(x^k)\|_2)$. We have $\|s^*\|_2 \geq K_0 \|\nabla J(x^k)\|_2$. Let $K_1 = (K_0/2) > 0$ and let $\tau = K_1 \|\nabla J(x^k)\|_2$. Then $\tau > 0$ since we only need to consider the case $\|\nabla J(x^k)\|_2 \neq 0$. Because $\{s(t)\}$ converges to s^* , there exists a t' such that $\|s^* - s(t)\|_2 < \tau$ for all $t > t'$. Consequently, $\|s(t)\|_2 > \|s^*\|_2 - \tau \geq K_0 \|\nabla J(x^k)\|_2 - K_1 \|\nabla J(x^k)\|_2 = K_1 \|\nabla J(x^k)\|_2$ for all $t > t'$. \square

Proof of Theorem 2.1: From Corollary 2.1, we see that $\forall k$, with $\nabla J(x^k) \neq 0$, there must exist a t' such that the stopping criteria (16)

$$\|s(t)\|_2 > K_1 \|\nabla J(x^k)\|_2, \quad \text{for some } K_1 > 0, \quad \forall t > t' \quad (A7)$$

is satisfied. Furthermore, by assumption, we have $\frac{1}{2} s(t)^T C(x^k) s(t) - s(t)^T K_2 s(t) \geq 0, \forall t$; and from (A5), we have $-\frac{1}{2} s(t)^T C(x^k) s(t) - \nabla J(x^k)^T s(t) \geq 0, \forall t$; these two inequalities lead to

$$-K_2 \|s(t)\|_2^2 \geq \nabla J(x^k)^T s(t), \quad \forall t. \quad (A8)$$

Then, the following proof mostly follows the proof of the convergence theorem for descent algorithms given in [1, pp. 203–204]. A Descent Lemma given in [1, pp. 203–204] states that if $J(\cdot)$ is continuously differentiable and there exists a positive constant K such that $\|\nabla J(x) - \nabla J(y)\| \leq K \|x - y\|, \forall x, y \in \mathbb{R}^n$, then $J(x+y) \leq J(x) + \nabla J(x)^T y + (K/2) \|y\|_2^2, \forall x, y \in \mathbb{R}^n$. By the assumptions on $J(\cdot)$ given in Section I, we may use the Descent Lemma and (A8) to obtain

$$\begin{aligned} J(x^k + \gamma^k s^k) & \leq J(x^k) + \gamma^k \nabla J(x^k)^T s^k + \frac{K}{2} \gamma^{k^2} \|s^k\|_2^2 \\ & \leq J(x^k) - \gamma^k \left(K_2 - \frac{K \gamma^k}{2} \right) \|s^k\|_2^2. \end{aligned} \quad (A9)$$

If $0 < \gamma^k < (K_2/K)$, we can derive that

$$-\gamma^k \left(K_2 - \frac{K \gamma^k}{2} \right) \leq -\frac{K_2}{2} \gamma^k. \quad (A10)$$

By (A9) and (A10), we have

$$J(x^k + \gamma^k s^k) \leq J(x^k) - \frac{K_2}{2} \gamma^k \|s^k\|_2^2. \quad (A11)$$

If $\lambda \geq (K_2/K)$, $0 < \beta < 1$, since $\{\beta^m \lambda\}$, $m = 0, 1, \dots$ is a monotonic decreasing sequence approaching 0, there must exist an m such that $\beta^m \lambda < (K_2/K)$. Let m_k be the smallest nonnegative integer satisfying the above inequality, then the following holds

$$\beta \frac{K_2}{K} \leq \beta^{m_k} \lambda < \frac{K_2}{K}. \quad (A12)$$

If $0 < \lambda < (K_2/K)$, $0 < \beta < 1$, then

$$\beta \lambda < \lambda < \frac{K_2}{K}. \quad (A13)$$

In this case, we can view $m_k = 0$ which is the smallest nonnegative integer for $\beta^{m_k} \lambda < (K_2/K)$. Now, for any $\lambda > 0$, $0 < \beta < 1$, and let m_k be either the m_k determined by (A12) or $m_k = 0$ by (A13); from (A11), there must exist an $m'_k \leq m_k$ (or $m'_k = 0$ if $m_k = 0$) such that

$$J(x^k + \beta^{m'_k} \lambda \hat{s}^k) \leq J(x^k) - \frac{K_2}{2} \beta^{m'_k} \lambda \|\hat{s}^k\|_2^2. \quad (\text{A14})$$

Note that $0 < \gamma^k < (K_2/K)$ is sufficient for (A14) to hold explains why $m'_k \leq m_k$. This shows that Step 8 of our algorithm will terminate for certain m'_k . Since $\beta^{m'_k} \geq \beta^{m_k}$, from (A12) and (A13), $\beta^{m'_k} \lambda \geq \min\{\beta \lambda, \beta(K_2/K)\}$. Let $\tau \equiv \frac{1}{2} K_2 \cdot \min\{\beta \lambda, \beta(K_2/K)\}$, then τ is finite and positive, and

$$J(x^k + \beta^{m'_k} \lambda \hat{s}^k) \leq J(x^k) - \tau \|\hat{s}^k\|_2^2. \quad (\text{A15})$$

Then, each iteration of our algorithm ensures a decrement of the objective function by at least the amount $\tau \|\hat{s}^k\|_2^2$. Since $J(x)$ is bounded from below, we assume $c \in \mathbb{R}$ is a lower bound of $J(x)$, then from (A15), we have

$$0 \leq J(x^{k+1}) - c \leq J(x^k) - c - \tau \|\hat{s}^k\|_2^2, \quad \forall k. \quad (\text{A16})$$

Then, by (A16), $\sum_{k=0}^{\infty} \|\hat{s}^k\|_2^2 \leq (J(x^0) - c/\tau) < \infty$, and (A7) shows that $\lim_{k \rightarrow \infty} \nabla J(x^k) = 0$. \square

REFERENCES

- [1] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. London: Prentice-Hall, 1989.
- [2] L. O. Chua and P. M. Lin, *Computer Aided Analysis of Electronic Circuits*. Englewood Cliffs, NJ: Prentice-Hall, 1975.
- [3] W. F. Tinney and J. W. Walker, "Direct solutions of sparse network equations by optimally ordered triangular factorization," *Proc. IEEE*, vol. 55, no. 11, pp. 1801-1809, Nov. 1967.
- [4] C.-H. Lin, "A new parallel processing algorithm for power system state estimation problems," master's thesis, Dept. of Elec. Engr., Nat'l Tsing Hua Univ., Hsinchu, Taiwan, 1991 (in Chinese).
- [5] F. C. Schweppe and J. Wildes, "Power system static-state estimation, part I: exact model," *IEEE Trans. Power App. Syst.*, vol. PAS-89, no. 1, pp. 120-125, Jan. 1970.
- [6] F. C. Schweppe, "Power system static-state estimation, part III: implementation," *IEEE Trans. Power App. Syst.*, vol. PAS-89, no. 1, pp. 130-135, Jan. 1970.
- [7] A. Simoes-Costa and V. H. Quintana, "A robust numerical technique for power system state estimation," *IEEE Trans. Power App. Syst.*, vol. PAS-100, pp. 691-698, Feb. 1981.
- [8] A. Garcia, A. Monticelli, and P. Abreu, "Fast decoupled state estimation and bad data processing," *IEEE Trans. Power App. Syst.*, vol. PAS-98, pp. 1645-1652, Sep. 1979.
- [9] S.-Y. Lin, C.-H. Lin, and S.-L. Yu, "An efficient descent algorithm for a class of unconstrained optimization problems of nonlinear large mesh-interconnected systems," in *Proc. 32nd IEEE Conf. Dec. & Contr.*, Dec. 1993.
- [10] S.-Y. Lin and C.-H. Lin, "An implementable distributed state estimator and bad data processing schemes for electric power systems," *IEEE Trans. Power Syst.*, vol. 9, no. 3, pp. 1277-1284, Aug. 1994.
- [11] S.-Y. Lin, "A parallel processing multi-coordinate descent method with line search - algorithm and convergence," in *Proc. 30th IEEE Conf. Dec. Contr.*, Dec. 1991, pp. 2096-2097.
- [12] D. Luenberger, *Linear and Nonlinear Programming*, 2nd ed. Reading, MA: Addison-Wesley, 1984.

On the Possible Divergence of the Projection Algorithm

Erjen Lefeber and Jan Willem Polderman

Abstract—It is shown by means of an example that the projection algorithm does not always converge.

I. INTRODUCTION

It is well known that parameter identification of linear systems depends very much on the excitation of the signals. Generally speaking, all identification algorithms require the signals to be sufficiently exciting. In applications such as adaptive control, however, excitation is often not possible. The question then arises how useful the standard identification schemes are. In this note we consider the case where the data can be modeled exactly by a linear time invariant discrete-time model. It is a fact, that for such systems recursive least squares always produce a convergent sequence of parameter estimates, although it is of course not guaranteed that the limit will be the true parameter [1].

For the projection algorithm a similar result or its negation is to the best of our knowledge not available in the literature. Properties that can be derived without any assumptions on the signals can be found in [1]. Nothing is said about convergence there (see also [2, Problem 12.14]). In [3], the algorithm is used for adaptive pole assignment. Since the adaptive algorithm could be analyzed without proving convergence of the parameter estimates, the possible convergence is not studied there either.

In this note we show by means of an example that the projection algorithm does not necessarily converge. This is in contrast with recursive least squares.

The construction of the counter example is as follows. Firstly we construct a sequence of real vectors that satisfies at least some of the properties of the projection algorithm and which does not converge. Secondly we show that the sequence could as well have been obtained by applying the projection algorithm to an appropriate input/output system. Hence, rather than fitting the estimates to the data, we fit the data to the estimates.

II. THE PROJECTION ALGORITHM

For the sake of completeness, we briefly describe the projection algorithm. Let the system be described by

$$y(k+1) = \bar{\theta}^T \phi(k) \quad \bar{\theta} \in \mathbb{R}^n. \quad (1)$$

The projection algorithm is defined as follows: Suppose that the estimate of $\bar{\theta}$ at time k is θ_k , define $G_{k+1} := \{\theta \in \mathbb{R}^n \mid y(k+1) = \theta^T \phi(k)\}$. Define θ_{k+1} as the orthogonal projection of θ_k on G_{k+1} . The recursion is given by

$$\theta_{k+1} = \theta_k + \frac{\phi(k)}{\|\phi(k)\|^2} (y(k+1) - \theta_k^T \phi(k)). \quad (2)$$

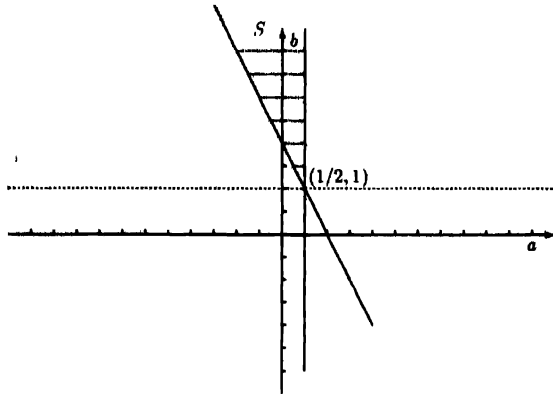
Notice that G_{k+1} contains the true parameter $\bar{\theta}$. Regardless of the input sequence, the following two properties hold.

- Property 2.1:** 1) For all k : $\|\bar{\theta} - \theta_{k+1}\| \leq \|\bar{\theta} - \theta_k\|$.
2) $\lim_{k \rightarrow \infty} (\theta_{k+1} - \theta_k) = 0$.

It is obvious that from Property 2.1 we cannot conclude that θ_k is a fundamental sequence, and in fact we will see that it need not be.

Manuscript received November 10, 1993.

The authors are with the Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
IEEE Log Number 9407235.

Fig. 1. The set S .

III. A COUNTEREXAMPLE

The idea of the counterexample is that we will first construct a sequence $(a_k, b_k) \in \mathbb{R}^2$ with the properties that:

- 1) (a_{k+1}, b_{k+1}) is obtained from (a_k, b_k) by orthogonally projecting the latter onto a line passing through a fixed point.
- 2) The sequence does not converge.

Notice that this sequence is constructed in a similar way as the sequence of estimates in the projection algorithm. Subsequently we will show that the particular sequence is equal to the sequence of estimates produced by applying the projection algorithm to a particular first-order system. That will establish the claim that the algorithm does not necessarily produce a convergent sequence of estimates. The key idea is that we fit the data to the estimates rather than the estimates to the data.

A. Construction of the Sequence

The sequence $\{a_k, b_k\}$ will be defined inductively

$$(a_0, b_0) := (1/2, 4). \quad (3)$$

Suppose now that (a_k, b_k) has been constructed. Let L_k be the line passing through $(1/2, 1)$ and (a_k, b_k) . Define (a_{k+1}, b_{k+1}) as the orthogonal projection on a line L_{k+1} yet to be defined. L_{k+1} will be a line passing through $(1/2, 1)$ with the property that the distance between (a_k, b_k) and its orthogonal projection on L_{k+1} is exactly $1/(k+1)$. There are two possibilities for L_{k+1} , one which requires a clockwise rotation of L_k to obtain L_{k+1} and one for which this rotation would be counter clockwise. This freedom of choice will now be used as follows. Define the region $S := \{(a, b) \mid -1 < -2a < b - 2 \wedge b > 1\}$. See Fig. 1. Determine the two possibilities for (a_{k+1}, b_{k+1}) . When both points are in S , rotate L_k in the same direction as L_{k-1} was rotated to obtain L_k , to get L_{k+1} , otherwise rotate L_k in the opposite direction. L_1 of course requires a counter clockwise rotation of L_0 . Notice that now every $(a_k, b_k) \in S$. Of course, the recursion could in principle be written in formulas; we feel, however, that this would not add much to our understanding.

Lemma 3.1: i) The sequence $\{(a_k, b_k)\}$ is well-defined. ii) The sequence $\{(a_k, b_k)\}$ does not converge.

Proof:

- i) Define $r_k := \sqrt{(1/2 - a_k)^2 + (1 - b_k)^2}$ and $\delta_{k+1} := \sqrt{(a_{k+1} - a_k)^2 + (b_{k+1} - b_k)^2}$. From the construction it follows that

$$r_{k+1}^2 + \delta_{k+1}^2 = r_k^2. \quad (4)$$

Since $\delta_{k+1} = 1/(k+1)$ it follows that

$$r_{k+1}^2 = r_k^2 - 1/(k+1)^2. \quad (5)$$

If we disregard for a moment the restriction imposed by S , we conclude that (a_{k+1}, b_{k+1}) can be constructed from (a_k, b_k) provided $r_k^2 - 1/(k+1)^2 > 0$. Now, from (5) it follows that

$$r_{k+1}^2 = r_0^2 - \sum_{j=0}^k 1/(j+1)^2 \quad (6)$$

hence we should have $r_0^2 > \pi^2/6$, since in our case $r_0^2 = 9$ this condition is satisfied.

As a byproduct we obtain that $\lim_{k \rightarrow \infty} r_k^2 = 9 - \pi^2/6$. It should be clear that from this we can also conclude that the requirement that $(a_k, b_k) \in S$ does not impose a restriction on the existence of the sequence.

- ii) From the fact that $r_k \rightarrow \sqrt{9 - \pi^2/6}$ and since $\delta_{k+1} = 1/(k+1)$, it follows that $\angle(L_k, L_{k+1})$ is $O(1/k+1)$. Therefore the sequence of lines $\{L_k\}$ does not converge and hence nor does $\{(a_k, b_k)\}$. \square

Lemma 3.2. Consider the i/o system

$$y(k+1) = (1/2)y(k) + u(k), \quad y(0) = 1.$$

There exists an input sequence $\{u(k)\}$, such that the projection algorithm, initialized in (a_0, b_0) generates $\{(a_k, b_k)\}$ as the sequence of estimates.

Proof: This is now easy. All we have to do is make sure that at time $k+1$, $G_{k+1} = L_{k+1}$, or equivalently, $(a_{k+1}, b_{k+1}) \in G_{k+1}$ and $y(k) \neq 0$. Otherwise stated $u(k)$ has to be such that

$$1/2 y(k) + u(k) = a_{k+1} y(k) + b_{k+1} u(k). \quad (7)$$

Hence we should take

$$u(k) = \frac{a_{k+1} - 1/2}{1 - b_{k+1}} y(k). \quad (8)$$

Since $(a_{k+1}, b_{k+1}) \in S$, this can indeed be done.

To complete the proof we have to check that for all k the output $y(k)$ will be nonzero. From (8) it follows that

$$y(k+1) = \left(1/2 + \frac{a_{k+1} - 1/2}{1 - b_{k+1}}\right) y(k). \quad (9)$$

Since $y(0) = 1$, and since $(a_{k+1}, b_{k+1}) \in S$, it follows from (9) that $y(k) \neq 0$.

Notice that since $(a_k, b_k) \in S$, we actually have that the sequences u and y are bounded. \square

We have now proved the following theorem.

Theorem 3.3: There exists a system of the form (1), a bounded input sequence u and an initialization of the projection algorithm, such that the resulting sequence of estimates does not converge.

IV. CONCLUSION

By means of an example, we have shown that the sequence of estimates generated by the projection algorithm does not necessarily converge. Of course, the sequence of inputs needed for the example is fairly artificial. In applications such as adaptive control, however, it is most desirable to derive as many properties of the identification part as possible without having to rely on the specific nature of the input. For the input will depend in a highly nonlinear fashion on the estimates. Our construction shows that convergence is not automatically among the properties that can be derived without additional assumptions on the input sequence.

REFERENCES

- [1] G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
- [2] G. C. Goodwin and R. H. Middleton, *Digital Control and Estimation*. Englewood Cliffs, NJ: Prentice-Hall, 1990.
- [3] J. W. Polderman, "A state space approach to the problem of adaptive pole assignment," *Math. Contr., Signals, Syst.*, vol. 2, pp. 71-94, 1989.

A Comment on the Method of the Closest Unstable Equilibrium Point in Nonlinear Stability Analysis

E. Noldus and M. Loccupier

Abstract—A counterexample is presented to a theorem which has been proposed as a theoretical basis for the method of the closest unstable equilibrium point to estimate asymptotic stability regions in nonlinear systems. An additional condition is formulated under which the theorem is valid. Its implications on the applicability of the method are discussed.

I. INTRODUCTION

The method of the closest unstable equilibrium point (c.u.e.p.) is a well-known direct method of the Lyapunov type for estimating regions of asymptotic stability (RAS) in nonlinear systems analysis. The method has been described, among others, by Chiang *et al.* [1], [2] and various applications, for example to the power system transient stability problem have been reported [3]–[5]. Its basic principle is the following: Consider an autonomous nonlinear dynamical system

$$\dot{x} = f(x) \quad (1)$$

where $x \in R^n$ represents the state and $f(\cdot)$ satisfies the sufficient conditions for the existence and the uniqueness of the solutions for given initial conditions. Suppose that a scalar function $V(x) \in C^n$, $n \geq 1$, can be found such that along the solutions of (1)

$$\begin{aligned} \dot{V}(x) &\leq 0, \forall x \in R^n \\ &= 0 \Leftrightarrow \dot{x} = 0. \end{aligned} \quad (2)$$

By (2), $V(x)$ is a Lyapunov function of (1) in R^n . Let \bar{x}_* be a locally asymptotically stable (l.a.s.) equilibrium state and let $\Omega(\bar{x}_*) \subset R^n$ be its exact RAS. Suppose that on the stability boundary $\partial\Omega(\bar{x}_*)$, $V(x)$ reaches an absolute minimum at $x = \hat{x}_*$ and let

$$\min_{x \in \partial\Omega(\bar{x}_*)} V(x) = V(\hat{x}_*) = V_{\min}. \quad (3)$$

Then it is well known that \hat{x}_* is an unstable equilibrium point of the system (1) [1]. Furthermore for any k in the interval

$$V(\hat{x}_*) < k \leq V_{\min}$$

the set

$$S \triangleq \{x: V(x) < k\}$$

the union of a number of connected, disjoint subsets

$$S = S_1 \cup S_2 \cup \dots \cup S_i;$$

$$S_i \cap S_j = \emptyset \quad \text{for } i \neq j$$

Manuscript received December 20, 1993; revised April 11, 1994.
The authors are with the University of Ghent, Department of Control Engineering and Automation, Technologiepark-Zwijnaarde 9, B-9052 Gent, Belgium.

EE Log Number 9407220.

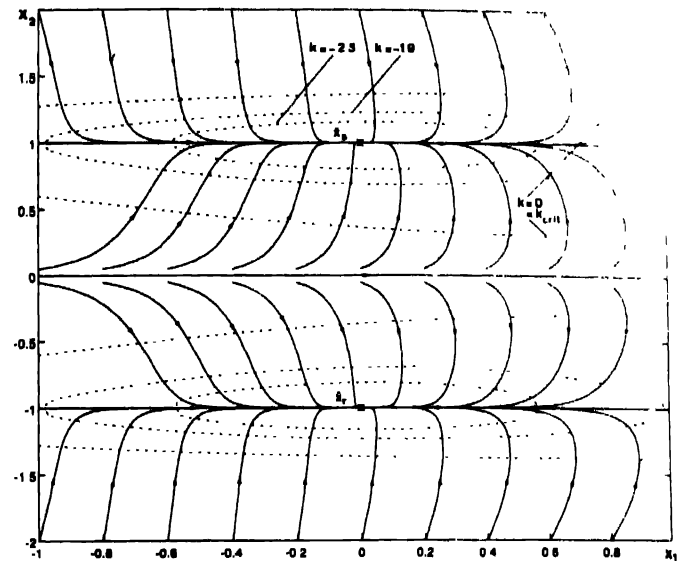


Fig. 1. Phase portrait of the system (5), (6) and level sets ∂S_1 for varying level values k , for the numerical values $a = 2$, $b = 1$, $\mu = 0.8$.

one of which, say S_1 contains \hat{x}_* . This subset S_1 is a RAS for \hat{x}_*

$$S_1 \subseteq \Omega(\hat{x}_*).$$

The largest stability region S_1 is obtained for $k = V_{\min}$. In [1] Chiang and Thorp have reported a theorem pertaining to the existence of the minimum V_{\min} , and a scheme for computing the corresponding stability region S_1 based on it.

Theorem [1]: If system (1) has a Lyapunov function $V(x)$ in R^n which satisfies (2) and if $\Omega(\hat{x}_*)$ is not dense in R^n , then V_{\min} as defined by (3) exists and \hat{x}_* is an unstable equilibrium state.

The proof relies on the property that if for $k = q$ the set \bar{S}_1 is a closed and bounded neighborhood of \hat{x}_* which contains no other equilibria, and if for some $p > q$ there are no equilibrium states in the set $\bar{S}_1|_{k=p} - \bar{S}_1|_{k=q}$ then

$$\bar{S}_1|_{k=p} \text{ is also closed and bounded.} \quad (4)$$

In Section II a counterexample to this result and to the property (4) is presented. It is pointed out, however, that the theorem is valid under the additional assumption that all trajectories on the stability boundary $\partial\Omega(\hat{x}_*)$ are bounded for $t \geq 0$. Section III discusses the implications of this proposition for the c.u.e.p. method.

II. EXAMPLE

Consider an example of the form

$$\dot{x} = f(x) \triangleq -\frac{\partial V(x)}{\partial x} \quad (5)$$

where $x \in R^2$ and

$$V(x) \triangleq e^{-x_1} - (2x_2^2 - x_2^4)v_1(x_1) \quad (6)$$

with

$$v_1(x_1) = [e^{-x_1} + ae^{-\mu x_1^2} + b]$$

and $a > 0$, $b > 0$ and $\mu > 0$. Then

$$\dot{V}(x) = \left[\frac{\partial V(x)}{\partial x} \right]^T \dot{x} = -\dot{x}^T \dot{x} \quad (7)$$

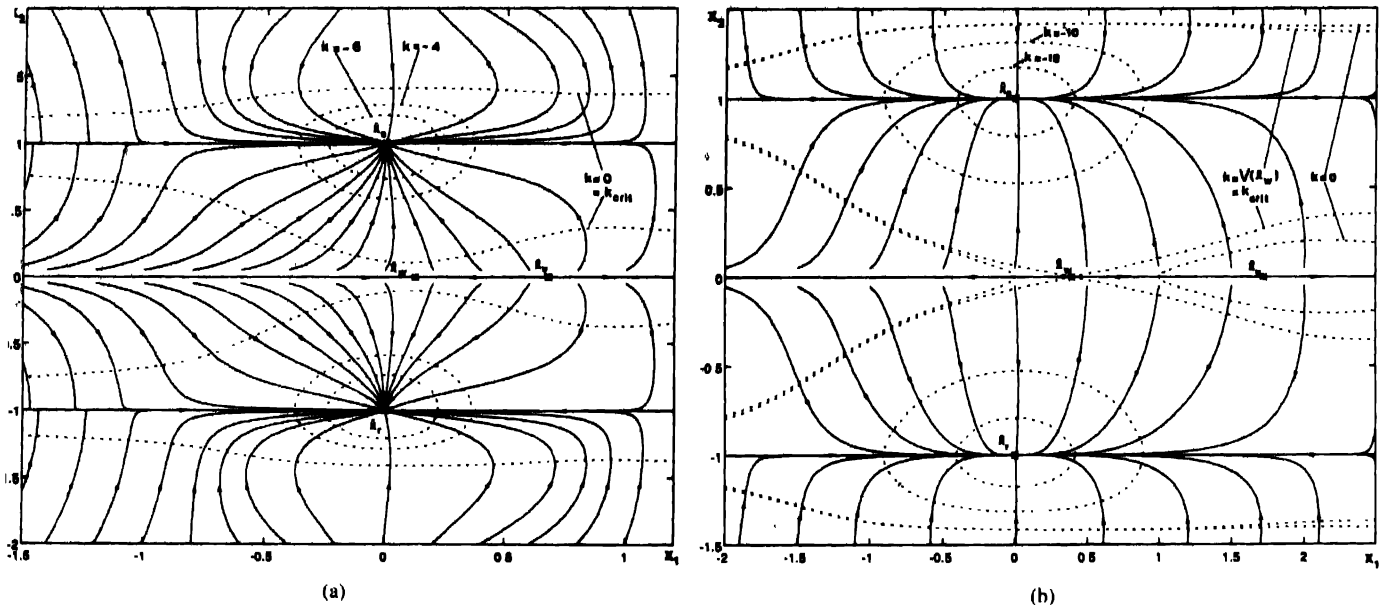


Fig. 2. Phase portrait and level sets ∂S_1 for the system (5), (8) and a): for the numerical values $a = 8$, $b = 1$, $c = 0.75$, $\mu = 5$ (V_{\min} does not exist.); b): $a = 20$, $b = 1$, $c = 1$, $\mu = 1$ ($V_{\min} = V(\hat{x}_m)$).

such that $V(x)$ is a Lyapunov function in R^2 which satisfies the conditions (2). Differentiating $V(x)$ w.r.t. x_1 and x_2 yields the state equations

$$\begin{aligned}\dot{x}_1 &= e^{-x_1} - (2x_2^2 - x_2^4) \\ &\quad \cdot [e^{-x_1} + 2\mu a x_1 e^{-\mu x_1^2}] \\ \dot{x}_2 &= 4x_2(1 - x_2^2) \\ &\quad \cdot [e^{-x_1} + a e^{-\mu x_1^2} + b].\end{aligned}$$

The phase portrait is symmetric w.r.t. the x_1 -axis. There are two equilibrium states, $\hat{x}_s = (0, 1)$ and $\hat{x}_r = (0, -1)$. Computing the Jacobian $J(x) = \partial f(x)/\partial x$ and substituting the coordinates of the equilibria produces $J(\hat{x}_s) = J(\hat{x}_r) = \text{diag} [-2\mu a, -8(1 + a + b)]$, such that \hat{x}_s and \hat{x}_r are stable nodes. Furthermore

$$\begin{aligned}x_2 = 0 &\Rightarrow \dot{x}_2 = 0; \\ \dot{x}_1 &= e^{-x_1} \\ x_2 = \pm 1 &\Rightarrow \dot{x}_2 = 0; \\ \dot{x}_1 &= -2\mu a x_1 e^{-\mu x_1^2}\end{aligned}$$

such that there is a single trajectory coinciding with the x_1 -axis, while the straight lines $x_2 = 1$ and $x_2 = -1$ each consist of two trajectories converging to, respectively, \hat{x}_s and \hat{x}_r for $t \rightarrow +\infty$. Finally all solutions are bounded for $t \geq 0$, except the solution whose trajectory coincides with the x_1 -axis (see Appendix A). Because of (7) all solutions which are bounded for $t \geq 0$ converge to an equilibrium state as $t \rightarrow +\infty$. Fig. 1 displays the system's phase portrait. Replacing (6) by

$$V(x) \triangleq e^{-x_1} - c e^{-\mu x_1^2} - (2x_2^2 - x_2^4)v_2(x_1) \quad (8)$$

where $a > 0$, $b > 0$, $c > 0$, $\mu > 0$ and

$$v_2(x_1) = [e^{-x_1} + (a - c)e^{-\mu x_1^2} + b]$$

produces a slightly modified example with state equations

$$\begin{cases} \dot{x}_1 = e^{-x_1} - 2\mu c x_1 e^{-\mu x_1^2} + (2x_2^2 - x_2^4) \frac{dv_2(x_1)}{dx_1} \\ \dot{x}_2 = 4x_2(1 - x_2^2)v_2(x_1) \end{cases} \quad (9)$$

where $dv_2(x_1)/dx_1 = -[e^{-x_1} + 2\mu(a - c)x_1 e^{-\mu x_1^2}]$.

If $a \geq c$ and c is sufficiently large, then the structure of the phase portrait is identical to the previous one, except that there are two additional equilibrium states $\hat{x}_m(p, 0)$ and $\hat{x}_n(q, 0)$ on the positive x_1 -axis. It is easily verified that \hat{x}_m (closest to the origin) is a saddle point and \hat{x}_n is an unstable node. Now there are three trajectories on the x_1 -axis (see Fig. 2(a) and (b)). In both examples $\Omega(\hat{x}_s) = \{x; x_2 > 0\}$ and $\Omega(\hat{x}_r) = \{x; x_2 < 0\}$ while $\partial\Omega(\hat{x}_s) = \partial\Omega(\hat{x}_r) = \{x; x_2 = 0\}$.

For the first example Fig. 1 shows the set S_1 for increasing values of $k > V(\hat{x}_s) = -(a + b)$. Note that $\lim_{x_1 \rightarrow -\infty} V(x) \triangleq V_1(x_2) = -b(2x_2^2 - x_2^4)$, which reaches an absolute minimum $V_{\min}(x_2) = -b$ at $x_2 = \pm 1$, and a relative maximum $V_{\max}(x_2) = 0$ at $x_2 = 0$.

Recall that the stability boundary $\partial\Omega(\hat{x}_s) = \partial\Omega(\hat{x}_r)$ coincides with the axis $x_2 = 0$. Hence S_1 remains bounded for $-(a + b) < k < -b$, but it becomes unbounded for $-b \leq k \leq 0$. It remains a region of attraction for \hat{x}_s as long as $k \leq 0$. The unboundedness of S_1 for $-b \leq k \leq 0$ contradicts (4). Although the system has a Lyapunov function in R^2 which satisfies (2) and $\Omega(\hat{x}_s)$ is not dense in R^2 , $V(x)$ does not have a minimum on $\partial\Omega(\hat{x}_s)$, which is counterexample to Chiang's [1] theorem quoted above.

III. DISCUSSION

In the procedure sketched in the introduction the set S_1 grows monotonically and continuously for increasing $k > V(\hat{x}_s)$, since $V(x) \in C^1$ where $r \geq 1$. S_1 remains a region of attraction of \hat{x}_s as long as ∂S_1 does not intersect the stability boundary $\partial\Omega(\hat{x}_s)$. Let

$$V_0 \triangleq \inf_{x \in \partial\Omega(\hat{x}_s)} V(x). \quad (10)$$

Then the largest obtainable stability region of the form S_1 corresponds to the critical value $k_{crit} = V_0$. In the first example $V_0 = 0$. If $V(x)$ possesses an absolute minimum V_{\min} for $x \in \partial\Omega(\hat{x}_s)$, then $V_0 = V_{\min}$. The existence of a minimum is guaranteed if

$$\text{All trajectories on } \partial\Omega(\hat{x}_s) \text{ are bounded for } t \geq 0 \quad (11)$$

and hence by (2) converge to an equilibrium state as $t \rightarrow +\infty$ (see Appendix B). So (11) constitutes an additional (sufficient) condition which guarantees the applicability of the c.u.e.p. method. Property (11) certainly holds if it can be shown that $\Omega(\hat{x}_s)$ is bounded, or if it can be shown that all trajectories are bounded

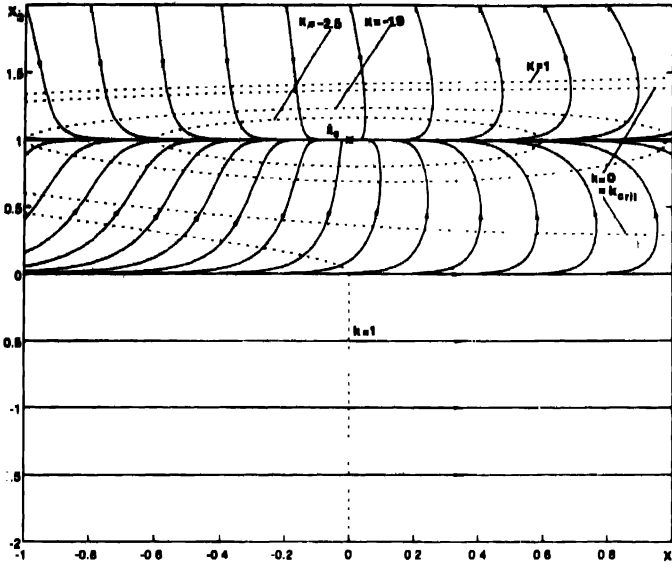


Fig. 3. Phase portrait and level sets ∂S_1 for the system (5), (8) and for the numerical values $a = 2$, $b = 1$, $\mu = 0.8$.

for $t \geq 0$. The latter property can often be proved using a suitable Lyapunov function. For example if $V(x)$ satisfies (2) and is radially unbounded, then all solutions are bounded for $t \geq 0$. In [6], [7] the structure of the stability boundary has been studied under a set of assumptions including the convergence of all solutions on $\partial\Omega(\hat{x}_*)$ to an equilibrium state.

The second example shows that even when there are equilibrium states on $\partial\Omega(x_*)$, V_{\min} does not necessarily exist. If

$$V(x_n) = e^{-\mu} - e^{-\mu\rho^2} > \lim_{t \rightarrow +\infty} V(x) = 0$$

then V_{\min} does not exist, while $V_0 = 0$. This case is illustrated in Fig. 2(a). If $V(\hat{x}_n) < 0$ then V_{\min} exists and the c.u.e.p. method can be applied with $k_{crit} = V(\hat{x}_n)$ (see Fig. 2(b)).

If V_{\min} does not exist then in general $k_{crit} = V_0$ cannot be determined since in (10), $\partial\Omega(x_*)$ is unknown. In such cases the only conclusion possible is that for increasing k the set S_1 remains a RAS as long as it is bounded and does not contain other equilibria than x_* . To illustrate the latter point consider a final example of the form (5) where

$$V(x) \triangleq e^{-x_1} - \varphi(x_2)[e^{-x_1} + ae^{-\mu x_1^2} + b] \quad (12)$$

and

$$\begin{aligned} \varphi(x_2) &\triangleq 2x_2^2 - x_2^4 \quad \text{for } x_2 \geq 0 \\ &\triangleq 0 \quad \text{for } x_2 < 0. \end{aligned}$$

Note that $V(x) \in C^1$. This system's phase portrait is identical to the phase portrait of Fig. 1 in the half plane $x_2 \geq 0$, while for $x_2 < 0$ all trajectories are straight lines parallel to the x_1 -axis ($x_2 = \text{constant}$; $x_1 \rightarrow +\infty$ for $t \rightarrow +\infty$). The system has a single equilibrium state $\hat{x}_*(0, 1)$ with stability region $\Omega(\hat{x}_*) = \{x: x_2 > 0\}$.

For increasing k the set S_1 remains a RAS of \hat{x}_* as long as it is bounded. For $k > 0$, S_1 is unbounded and contains trajectories that converge to infinity for $t \rightarrow +\infty$, hence it is not a RAS any more (see Fig. 3). The critical value $k_{crit} = 0$ cannot be determined if the stability boundary $\partial\Omega(\hat{x}_*)$ is unknown.

Finally it should be stressed that (11) is only a sufficient but not a necessary condition for the existence of an absolute minimum of $V(x)$ for $x \in \partial\Omega(\hat{x}_*)$. Any other sufficient condition for the existence of the minimum yields a valid alternative version of the fundamental theorem of the c.u.e.p. method. In [2, Theorem 9], and

in [8, Theorem 4.1], Chiang has provided such an alternative which is suitable for studying the power system transient stability problem. An energy function $V(x)$ is introduced, which for the application to the power system model plays the role of a Lyapunov function, upon which three conditions are imposed:

- 1) Along any trajectory $x(t)$, $\dot{V}(x(t)) \leq 0$.
- 2) Along any nontrivial trajectory the set $\{t \in R; \dot{V}(x(t)) = 0\}$ has measure zero in R .
- 3) Along any trajectory, if $V(x(t))$ is bounded then $x(t)$ is bounded.

The conditions of Theorem 4.1 in [8] guarantee the existence of an absolute minimum of the energy function on the stability boundary. In fact Chiang has shown that under these conditions the c.u.e. point, where the absolute minimum is reached, is generically unique.

APPENDIX

A. Proof of the Boundedness of Solutions

Since the Lyapunov function (6) is a special case of (8) consider the system defined by (5), (8). Let

$$W(x) \triangleq (1 - x_2^2)^2.$$

Then using (9)

$$\begin{aligned} \dot{W}(x) &= -16x_2^2(1 - x_2^2)^2 v_2(x_1) \\ &\leq -16x_2^2(1 - x_2^2)^2 b \leq 0 \end{aligned} \quad (13)$$

since $a \geq c$, (13) implies that

$$\min[x_2^2(0), 1] \leq x_2^2(t) \leq \max[x_2^2(0), 1] \quad (14)$$

$\forall t \geq 0$ such that

$$\dot{W}(x) \leq -\lambda_1 W(x)$$

hence

$$W(x) \leq [1 - x_2^2(0)]^2 e^{-\lambda_1 t}; \quad \forall t \geq 0 \quad (15)$$

where $\lambda_1 \triangleq 16b \min[x_2^2(0), 1] > 0$ if $x_2(0) \neq 0$. Furthermore by (9)

$$\dot{x}_1 = (1 - x_2^2)^2 [e^{-x_1} + 2\mu(a - c)x_1 e^{-\mu x_1^2}] - 2\mu a x_1 e^{-\mu x_1^2} \quad (16)$$

hence

$$\dot{x}_1 > 0 \quad \text{for } x_1 \leq -m \text{ and } m \text{ sufficiently large.} \quad (17)$$

Finally by (16), $x_1 = 0$ and $x_2^2 \neq 1$ imply that $\dot{x}_1 > 0$, such that $x_1(0) > 0$ implies that $x_1(t) > 0$ for all $t \geq 0$. Now by (16)

$$\dot{x}_1 \leq (1 - x_2^2)^2 (1 + n) e^{-x_1} \quad \text{for } n > 0 \text{ and sufficiently large}$$

or, defining $\rho \triangleq (1 + n)[1 - x_2^2(0)]^2$, $y \triangleq e^{-x_1}$ and using (15)

$$\dot{y} \leq \rho e^{-\lambda_1 t}.$$

Integrating the latter expression yields

$$y(t) \leq y_0 - \frac{\rho}{\lambda_1} e^{-\lambda_1 t} < y_0; \quad \forall t \geq 0$$

and for some finite y_0 . It follows that

$$x_1(t) \leq \ln y_0 < \infty; \quad \forall t \geq 0 \quad (18)$$

if $x_1(0) > 0$ and $x_2(0) \neq 0$. (14), (17) and (18) prove the boundedness for $t \geq 0$ of all solutions for which $x_2(0) \neq 0$.

B. Existence of V_{\min}

Let $\hat{X} \triangleq \{\hat{x}_1, \dots, \hat{x}_m\}$ be the set of equilibrium points on $\partial\Omega(\hat{x}_*)$ and let $V_{\min} \triangleq \min V(\hat{x}_i)$ for $i = 1 \dots m$. If $x_0 \in \partial\Omega(\hat{x}_*)$ then the trajectory starting at x_0 at $t = 0$ converges to an equilibrium point $\hat{x}_i \in \partial\Omega(\hat{x}_*)$. Since along this trajectory $V(x)$ is nonincreasing it follows that $V(x_0) \geq V(\hat{x}_i) \geq V_{\min}$ for all $x_0 \in \partial\Omega(\hat{x}_*)$. This proves the existence of the minimum of $V(x)$ on $\partial\Omega(\hat{x}_*)$.

REFERENCES

- [1] H. D. Chiang and J. S. Thorp, "Stability regions of nonlinear dynamical systems: A constructive methodology," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 1229-1241, Dec. 1989.
- [2] H. D. Chiang, F. F. Wu, and P. P. Varaiya, "Foundations of direct methods for power system transient stability analysis," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 160-173, Feb. 1987.
- [3] N. A. Tsolas, A. Arapostathis, and P. P. Varaiya, "A structure preserving energy function for power system transient stability analysis," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 1041-1048, Oct. 1985.
- [4] J. L. Willems, "Direct methods for transient stability studies in power system analysis," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 332-341, Aug. 1971.
- [5] C. J. Tavora and O. J. M. Smith, "Stability analysis of power systems," *IEEE Trans. Power Apparatus Syst.*, vol. PAS-91, pp. 1138-1144, May/June 1972.
- [6] H. D. Chiang, M. W. Hirsch, and F. F. Wu, "Stability regions of nonlinear autonomous dynamical systems," *IEEE Trans. Automat. Contr.*, vol. AC-33, pp. 16-27, Jan. 1988.
- [7] J. Zaborszky, G. Huang, B. Zheng, and T. Leung, "On the phase portrait of a class of large nonlinear dynamic systems such as the power system," *IEEE Trans. Automat. Contr.*, vol. AC-33, pp. 4-15, Jan. 1988.
- [8] H. D. Chiang, "Analytical results on direct methods for power system transient stability analysis," *Advances in Control and Dynamic Systems XL*. New York: Academic, vol. 43, no. 3, pp. 275-334, 1991.

Boundaries of Conditional Quadratic Forms—A Comment on "Stabilization via Static Output Feedback"

D. Cheng and C. F. Martin

Abstract—Motivated by the above paper,¹ this note considers the boundaries of a quadratic form with all possible constraints over a given subspace. Essential upper (or lower) bounds are presented provided they exist. It mends a mild incompleteness in the proof of the main result.

1. INTRODUCTION

It is well known (see, e.g., [2]) that for a real symmetric matrix M

$$\min \sigma(M) \leq \frac{x^T M x}{x^T x} \leq \max \sigma(M), x \neq 0 \quad (1)$$

where $\sigma(M)$ is the set of eigenvalues of M .

In the above paper, necessary and sufficient conditions for the existence of a stabilizing static output feedback gain matrix were presented. In the proof of the main result, Theorem 3.1, the following fact was used (for the sake of consistency we use our notations).

Given a real symmetric matrix: $M_{n \times n}$, and a matrix $K_{m \times n}$ with $\text{rank}(K) = r < n$. Assume

$$x^T(M)x < 0; \forall x \in \text{Ker}(K). \quad (2)$$

Then

$$\alpha = \sup_{x \notin \text{Ker}(K)} \frac{x^T M x}{x^T K^T K x} < \infty, x \notin \text{Ker}(K). \quad (3)$$

Manuscript received January 26, 1994; revised June 28, 1994.

D. Cheng is with Gifford Fong Associates, Walnut Creek, CA 94596 USA.

C. F. Martin is with the Department of Mathematics, Texas Tech University, Lubbock, TX 79409 USA.

IEEE Log Number 9407224.

A mild incompleteness in the proof¹ is: They did not claim and prove that the α (defined in (3.4), which is the same as in (3)), is upper bounded, i.e., $\alpha < +\infty$. It is essential for constructing R ($R^{-1} \geq \alpha I$). It was pointed by a nominated reviewer that this boundedness is also assumed in [2] without proof. From the following discussion one sees that this fact is not trivial. We call it the problem of boundaries of conditional quadratic forms. It can be considered as a generalization of (1), because when $\dim(K) = 0$, our results in this note will coincide with (1).

In the next section we prove (3) by giving the essential upper bound, which may be of independent interest. Then in Section III we discuss all other constraints.

II. MAIN RESULT

Let the matrices $M_{n \times n}$, $K_{m \times n}$, with $\text{rank}(K) = r < n$, be as above. Then there exists a linear transformation Φ such that

$$K\Phi = (S \quad 0)$$

where S is the first r columns, and thus S has full rank. It follows that $S^T S$ is a positive definite matrix, so we can define a positive definite E as

$$E = (S^T S)^{1/2} > 0. \quad (4)$$

It is easy to prove that (2) is equivalent to the fact that after the linear transformation Φ , M has the following form

$$\Phi^T M \Phi = \begin{pmatrix} A & B \\ B^T & -Q \end{pmatrix}. \quad (5)$$

Then we define a characteristic matrix C as

$$C = E^{-1}(A + BQ^{-1}B^T)E^{-1}. \quad (6)$$

Using the above notations, we can prove the following theorem.

Theorem 1: Under condition (2), we have the essential upper bound as

$$\sup_{x \notin \text{Ker}(K)} \frac{x^T M x}{x^T K^T K x}, x \notin \text{Ker}(K) = \max \sigma(C). \quad (7)$$

Proof: It is clear that $x \in \text{Ker}(K)$, if and only if

$$y = \Phi^{-1}x = \begin{pmatrix} 0 \\ y_2 \end{pmatrix}.$$

Since Q is positive definite, $Q^{1/2} > 0$ is well defined. Using (5), a straightforward computation shows that

$$\begin{aligned} \sup_{x \notin \text{Ker}(K)} \frac{x^T M x}{x^T K^T K x}, x \notin \text{Ker}(K) &= \sup_y \frac{y^T \begin{pmatrix} A & B \\ B^T & -Q \end{pmatrix} y}{y^T \begin{pmatrix} S^T S & 0 \\ 0 & 0 \end{pmatrix} y} \\ &= \sup_y \frac{y_1^T (A + BQ^{-1}B^T)y_1 - \|Q^{1/2}y_2 - Q^{-1/2}B^T y_1\|^2}{y_1^T S^T S y_1} \\ &= \sup_{y_1 \neq 0} \frac{y_1^T (A + BQ^{-1}B^T)y_1}{y_1^T S^T S y_1} \end{aligned} \quad (8)$$

¹ A. Trofino-Neto and V. Kucera, *IEEE Trans. Automat. Contr.*, vol. 38, no. 5, pp. 764-765, May 1993.

where

$$y = \Phi^{-1}x = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad y_1 \neq 0.$$

The last equality is obtained by setting $y_2 = Q^{-1}B^T y_1$.

Recalling the definitions of E and C in (4) and (6), we get

$$\begin{aligned} & \sup_{y_1 \neq 0} \frac{y_1^T (A + BQ^{-1}B^T)y_1}{y_1^T S^T S y_1} \\ &= \sup_{Z \neq 0} \frac{Z^T C Z}{Z^T Z}, \text{ where } Z = E y_1. \end{aligned}$$

From (1), it can be seen that

$$\begin{aligned} & \sup_x \frac{x^T M x}{x^T K^T K x}, x \notin \text{Ker}(K) \\ &= \max \sigma(C) < +\infty. \end{aligned} \quad (9)$$

Since the transformation Φ is not unique, the last thing we have to do is to show that the essential upper bound obtained in (9) is independent of the choice of Φ . Theoretically, essential upper bound is unique. But, we use a particular Φ to get it, and the parameters A, B, C, E , etc. in the expression depend on Φ . So we must show that the upper bound is independent of Φ . Let $\tilde{\Phi} = (\tilde{\phi}_1 | \tilde{\phi}_2)$, where $\text{Span}(\phi_2) = K^\perp$, be another suitable linear transformation, and $\tilde{\Phi} = \Phi T = (\phi_1 | \phi_2)T$. Since $\text{Span}(\phi_2) = \text{Span}(\tilde{\phi}_2) = K^\perp$, it follows that

$$T = \begin{pmatrix} T_1 & 0 \\ T_2 & T_3 \end{pmatrix}.$$

Now, a straightforward computation shows that the corresponding expressions under the new transformation are

$$\begin{aligned} & \begin{pmatrix} S^T S & 0 \\ 0 & 0 \end{pmatrix} = (K\tilde{\Phi})^T (K\tilde{\Phi}) \\ &= T^T (K\Phi)^T (K\Phi) T = \begin{pmatrix} T_1^T S^T S T_1 & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned} & \begin{pmatrix} A & B \\ \tilde{B}^T & -\tilde{Q} \end{pmatrix} = \begin{pmatrix} T_1^T & T_2^T \\ 0 & T_3^T \end{pmatrix} \begin{pmatrix} A & B \\ B^T & -Q \end{pmatrix} \begin{pmatrix} T_1 & 0 \\ T_2 & T_3 \end{pmatrix} \\ &= \begin{pmatrix} T_1^T A T_1 + T_2^T B^T T_1 + T_1^T B T_2 - T_2^T Q T_2 & T_1^T B T_3 - T_2^T Q T_3 \\ T_1^T B^T T_1 - T_1^T Q T_2 & -T_1^T Q T_3 \end{pmatrix}. \end{aligned} \quad (10)$$

Using them, we finally have $\tilde{S}^T \tilde{S} = T_1^T S^T S T_1$, $\tilde{A} + \tilde{B} \tilde{Q}^{-1} \tilde{B}^T = T_1^T (A + BQ^{-1}B^T)$.

Therefore, the parameters obtained by the new transformation provide

$$\begin{aligned} & \sup_{y_1 \neq 0} \frac{y_1^T (A + \tilde{B} \tilde{Q}^{-1} \tilde{B}^T) y_1}{y_1^T \tilde{S}^T \tilde{S} y_1} \\ &= \sup_{Z \neq 0} \frac{Z^T C Z}{Z^T Z}, \text{ where } Z = E T_1 y_1. \end{aligned}$$

Q.E.D

III. GENERALIZATION

As we mentioned before, conditional quadratic form (3) is a generalization of the famous result (1). So the boundary problem of expression (3) has both theoretical and practical interests. To make (3) meaningful $x \notin \text{Ker}(K)$. Moreover, without constrain on $\text{Ker}(K)$ expression (3) has no boundary. (2) is a particular constrain. This section consider all other possible constraints on subspace $\text{Ker}(K)$. They may be applied to mini-max problems of quadratic forms.

Case 2: Condition (2) is replaced by

$$x^T (M)x > 0; \forall x \in \text{Ker}(K). \quad (11)$$

In this case, replace $-Q$ by Q in (5) and redefine C in (6) accordingly. Then a parallel discussion shows the following corollary.

Corollary 1: Under condition (11), we have

$$\sup_x \frac{x^T M x}{x^T K^T K x}, x \notin \text{Ker}(K) = \min \sigma(C). \quad (12)$$

Case 3: Assume now we have

$$x^T (M)x = 0; \forall x \in \text{Ker}(K). \quad (13)$$

In this case replace Q by zero in (5). Equation (8) becomes

$$\begin{aligned} & \sup_x \frac{x^T M x}{x^T K^T K x}, x \notin \text{Ker}(K) \\ &= \sup_y \frac{y_1^T (A) y_1 + 2 y_2^T B^T y_1}{y_1^T S^T S y_1}. \end{aligned} \quad (14)$$

From (14) one sees the following corollary.

Corollary 2: Under condition (13), if $B \neq 0$

$$\frac{x^T M x}{x^T K^T K x}, x \notin \text{Ker}(K)$$

has neither upper bound nor lower bound. If $B = 0$

$$\begin{aligned} \min \sigma(E^{-1} M E^{-1}) &\leq \frac{x^T M x}{x^T K^T K x} \leq \max \sigma(E^{-1} M E^{-1}), \\ x &\notin \text{Ker}(K). \end{aligned}$$

Proof: Observe (14). If $B = 0$, the conclusion follows from the standard result (1). If $B \neq 0$, choose

$$y_2 = \mu B^T y_1, \mu \in R.$$

Letting μ go to either $-\infty$ or $+\infty$, one sees that neither upper bound nor lower bound exist. Q.E.D.

Note that in Case 3 we have from (10) that $B = T_1^T B T_3$, and T_1 and T are nonsingular. It is clear that the conclusion in the above corollary is independent of the linear transformation.

Case 4. Assume

$$x^T (M)x \leq 0; \forall x \in \text{Ker}(K). \quad (15)$$

In this case the matrix Q in (5) is positive semi-definite. Then we have the following corollary.

Corollary 3: Under condition (15), if there exists a matrix H with suitable dimension such that $B = H Q$, then

$$\begin{aligned} \sup_x \left(\frac{x^T M x}{x^T K^T K x} \right) &= \max \sigma(E^{-1} (A + H Q H^T) E^{-1}), \\ x &\notin \text{Ker}(K). \end{aligned}$$

Otherwise

$$\frac{x^T M x}{x^T K^T K x}, x \notin \text{Ker}(K)$$

has neither upper bound nor lower bound.

Proof: Let $B = H Q$. Then we can choose

$$T = \begin{pmatrix} I & 0 \\ H^T & I \end{pmatrix}.$$

From (10), new M is block diagonal, and then (8) becomes

$$\begin{aligned} & \sup_x \frac{x^T M x}{x^T K^T K x}, x \notin \text{Ker}(K) \\ &= \sup_y \frac{y_1^T (A + H Q H^T) y_1 - y_2^T Q y_2}{y_1^T S^T S y_1} \\ &\leq \sup_{y_1} \frac{y_1^T (A + H Q H^T) y_1}{y_1^T S^T S y_1}. \end{aligned}$$

The conclusion follows.

If $B \neq HQ$, the rows of B are not all in $\text{Span row } Q$. Without loss of generality we assume

$$Q = \begin{pmatrix} Q_1 & 0 \\ 0 & 0 \end{pmatrix}, \text{ where } Q_1 > 0$$

and partition y_2 as

$$y_2 = \begin{pmatrix} y_{21} \\ y_{22} \end{pmatrix}.$$

Now (8) has the form

$$\frac{x^T M x}{x^T K^T K x}, \quad x \notin \text{Ker}(K) \\ \frac{y_1^T (A)y_1 + 2(y_{21}^T, y_{22}^T) B^T y_1 - t_{21}^T Q_1 y_{21}}{y_1^T S^T S y_1}.$$

The condition: "rows of B are not all in $\text{Span row } Q$ " implies that $3y_{22}$ is not always zero. It is obvious that this term can make the value of the fraction be both positive and negative unbounded.

One can also see from (10) that the condition $B = HQ$ is independent of the linear transformation.

Q.E.D.

Case 5: Assume

$$x^T (M)x \geq 0; \forall x \in \text{Ker}(K). \quad (19)$$

In this case replace $-Q$ by a positive semi-definite Q in (5). Similar to the proof of Case 4, we can show the following.

Corollary 4: Under condition (19), if there exists a suitable dimensional matrix H such that $B = HQ$, then

$$\inf_{x \notin \text{Ker}(K)} \left(\frac{x^T M x}{x^T K^T K x} \right) = \min \sigma(E^{-1}(A + HQH^T)E^{-1}),$$

Otherwise

$$\frac{x^T M x}{x^T K^T K x}, x \notin \text{Ker}(K)$$

has neither upper bound nor lower bound.

IV. CONCLUSION

In this note we discussed the problem of finding the boundaries of conditional quadratic forms with all possible constraints over a subspace. In all cases, the necessary and sufficient conditions for the existence of upper and/or lower bounds are presented. The essential bounds are obtained whenever they exist.

REFERENCES

- [1] R. Bellman, *Introduction to Matrix Analysis*. New York: McGraw-Hill, 1960.
- [2] J. C. Geromel, J. Bernussou, P. L. D. Peres, "Stabilizability of uncertain systems via linear programming," in *Proc. IEEE Conf. Decis. Contr.*, Austin, TX, 1988, pp. 1771-1775.

On Robust Stability of 2-D Discrete Systems

W.-S. Lu

Abstract—This note presents a study on robust stability of two-dimensional (2-D) discrete systems in the Fornasini-Marchesini (F-M) state space setting. A measure of stability robustness of a stable F-M model is defined. Relation of this measure to its counterpart in the Roesser state space and related computational issues are addressed. Three lower bounds of the stability-robustness measure defined are derived using an one-dimensional parameterization approach and a 2-D Lyapunov approach. A numerical example is included to illustrate the main results obtained.

1. INTRODUCTION

In this note, we present a study on robust stability of two-dimensional (2-D) discrete systems under unstructured perturbations. Throughout the concerned system is modeled in the Fornasini-Marchesini (F-M) local state space [1] as

$$x(i+1, j+1) = A_1 x(i, j+1) + A_2 x(i+1, j) \quad (1)$$

where $x(i, j) \in R^{n \times 1}$, $A_1, A_2 \in R^{n \times n}$. Recall that system (1) is asymptotically stable if and only if

$$\rho(z_1, z_2) \equiv \det(I_n - z_1 A_1 - z_2 A_2) \neq 0 \text{ for } (z_1, z_2) \in \bar{U} \quad (2)$$

where $\bar{U} = \{(z_1, z_2) : |z_1| \leq 1, |z_2| \leq 1\}$ [1]. To date, results on robust stability of 2-D discrete systems in a local state-space framework are only available for the Roesser model [2]–[6], and one might attribute this to the lack of a Lyapunov stability theory for the F-M model. The objectives of this note are twofold. First, we propose in Section II a quantitative measure, ν , for the unstructured, stable perturbations of a given stable 2-D F-M system, and derive in Section III a lower bound for ν . Issues on numerical evaluation of the bound obtained and the relation of the proposed stability robustness measure with its counterpart in the Roesser state space will also be addressed. Second, we propose in Section IV a Lyapunov approach to analyzing the robust stability of (1), leading to two lower bounds of ν . The proposed approach makes use of the 2-D Lyapunov equation [7] which is a generalization of the 2-D Lyapunov equation investigated recently by Hinamoto [8]. A numerical example is included in Section V to illustrate the main results of the paper.

In the rest of the paper we write $H > 0$ or $H \geq 0$ to mean that the symmetric matrix H is positive definite or positive semi-definite. For a real matrix $P \geq 0$, one can always write

$$P = U^{-1} \Sigma U$$

where U is orthogonal and $\Sigma = \text{diag}\{\sigma_1, \dots, \sigma_n\}$ with $\sigma_k \geq 0$, for $1 \leq k \leq n$. If we denote $\Sigma^{1/2} = \text{diag}\{\sigma_1^{1/2}, \dots, \sigma_n^{1/2}\}$, then

$$P = P^{1/2} P^{1/2}$$

where

$$P^{1/2} = \Sigma^{1/2} U \text{ and } P^{T/2} = (P^{1/2})^T.$$

Such a $P^{1/2}$ is called the nonsymmetric square root of P . The largest and smallest singular value of matrix H is denoted by $\tilde{\sigma}(H)$ and

Manuscript received September 1, 1993; revised February 28, 1994 and August 1, 1994. This work was supported in part by the Natural Sciences and Engineering Council of Canada and the Networks of Centres of Excellence program in Microelectronics (Micronet).

The author is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada V8W 3P6.

IEEE Log Number 9407223.

$\sigma(H)$, respectively, $\|H\|$ denotes the induced 2-norm of H , which is equal to $\bar{\sigma}(H)$. Throughout we shall use A to denote the $n \times 2n$ real matrix $[A_1 \ A_2]$ where A_1 and A_2 are from (1) and call it the system matrix of (1). A is said to be stable if (1) is stable i.e., A satisfies (2).

II. A MEASURE OF ROBUST STABILITY OF SYSTEM (1)

For asymptotically stable system (1), let $A = [A_1 \ A_2]$ and denote by S_u the set of all complex perturbations $\Delta A = [\Delta A_1 \ \Delta A_2] \in \mathbb{C}^{n \times 2n}$ with $A + \Delta A$ unstable, i.e.,

$$S_u = \{\Delta A : [\Delta A_1 \ \Delta A_2] \in \mathbb{C}^{n \times 2n}, A + \Delta A \text{ unstable}\}. \quad (3)$$

Two sets of matrices can be induced from set S_u as follows

$$S_{u1} = \{\Delta A_1 : \Delta A_1 \in \mathbb{C}^{n \times n}, [\Delta A_1 \ 0] \in S_u\} \quad (4)$$

$$S_{u2} = \{\Delta A_2 : \Delta A_2 \in \mathbb{C}^{n \times n}, [0 \ \Delta A_2] \in S_u\}. \quad (5)$$

In words, S_{u1} and S_{u2} are the collection of the first and the second $n \times n$ blocks, respectively, each of which itself causes system instability.

A measure for robust stability of system (1) is defined as

$$\nu = \inf_{\Delta A \in S_u} \|\Delta A\|. \quad (6)$$

Obviously, a perturbed system with system matrix $A + \Delta A$ is stable if $\|\Delta A\| < \nu$.

With the sets S_{u1} and S_{u2} defined in (4) and (5), two auxiliary quantities can be introduced as

$$\mu_1 = \inf_{\Delta A_1 \in S_{u1}} \|\Delta A_1\| \text{ and } \mu_2 = \inf_{\Delta A_2 \in S_{u2}} \|\Delta A_2\|. \quad (7)$$

Since $S_{u1} \subset S_u$ and $S_{u2} \subset S_u$, we have $\nu \leq \mu_1$, $\nu \leq \mu_2$, and hence

$$\nu \leq \min(\mu_1, \mu_2). \quad (8)$$

Shortly we shall see that a relation similar to (8) also holds when the system is modeled in Roesser state space.

Like the 1-D case, ν defined in (6) will be referred to as the stability radius for unstructured complex perturbations for state-space model (1). Since perturbations for a physical 2-D plant are always of real value, a more realistic robust stability measure can be defined as

$$\nu_r = \inf_{\Delta A \in T_u} \|\Delta A\|$$

where T_u is the set of all real perturbations $\Delta A = [\Delta A_1 \ \Delta A_2] \in \mathbb{R}^{n \times 2n}$ with $A + \Delta A$ unstable, i.e.,

$$T_u = \{\Delta A : [\Delta A_1 \ \Delta A_2] \in \mathbb{R}^{n \times 2n}, A + \Delta A \text{ unstable}\}.$$

Note that T_u is a proper subset of S_u and consequently the stability radius for unstructured real perturbations, ν_r , is always no less than ν . This means that ν_r is indeed a less conservative measure for stability robustness. For the 1-D case, the real stability radius has been investigated [17]–[22] and a complete solution has been obtained very recently by Qiu *et al.* [22]. The generalization of the approach of [22] to the 2-D case is however far from straightforward and seems adequate to leave it as a separate topic for research.

Concerning the relation of the proposed stability measure to that of Roesser model, we recall [1] that if a 2-D discrete system has been modeled in the Roesser state space as

$$\begin{bmatrix} x^h(i+1, j) \\ x^v(i, j+1) \end{bmatrix} = \begin{bmatrix} \hat{A}_1 & \hat{A}_2 \\ \hat{A}_3 & \hat{A}_4 \end{bmatrix} \begin{bmatrix} x^h(i, j) \\ x^v(i, j) \end{bmatrix} \equiv \hat{A} \begin{bmatrix} x^h(i, j) \\ x^v(i, j) \end{bmatrix} \quad (9a)$$

it can be remodeled in the Fornasini-Marchesini state-space as (1) with

$$x(i, j) = \begin{bmatrix} x^h(i, j) \\ x^v(i, j) \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 \\ \hat{A}_3 & \hat{A}_4 \end{bmatrix}, \text{ and } A_2 = \begin{bmatrix} \hat{A}_1 & \hat{A}_2 \\ 0 & 0 \end{bmatrix}. \quad (9b)$$

Assuming (9a) is stable, the commonly used stability robustness measure of the system (in the Roesser state space) is

$$\xi = \inf_{\Delta \hat{A} \in R_u} \|\Delta \hat{A}\|$$

where

$$R_u = \{\Delta \hat{A} : \hat{A} + \Delta \hat{A} \text{ unstable}\}$$

with

$$\Delta \hat{A} = \begin{bmatrix} \Delta \hat{A}_1 & \Delta \hat{A}_2 \\ \Delta \hat{A}_3 & \Delta \hat{A}_4 \end{bmatrix}.$$

According to the structures of A_1 and A_2 in (9b), perturbation matrix $\Delta \hat{A}$ in the Roesser model is the sum of perturbations for A_1 and A_2 in the F-M model, i.e.,

$$\Delta \hat{A} = \begin{bmatrix} 0 & 0 \\ \Delta \hat{A}_3 & \Delta \hat{A}_4 \end{bmatrix} + \begin{bmatrix} \Delta \hat{A}_1 & \Delta \hat{A}_2 \\ 0 & 0 \end{bmatrix} \equiv \Delta A_1 + \Delta A_2. \quad (10)$$

Hence

$$\xi = \inf_{\Delta \hat{A} \in R_u} \|\Delta \hat{A}\| = \inf_{\Delta A \in S_u} \|\Delta A_1 + \Delta A_2\|$$

where S_u is the set of unstable perturbations $\Delta A = [\Delta A_1 \ \Delta A_2]$ with $\Delta A_1, \Delta A_2$ defined by (10). From $S_{u1} \subset S_u$ and $S_{u2} \subset S_u$, it follows that

$$\begin{aligned} \xi &= \inf_{\Delta A \in S_u} \|\Delta A_1 + \Delta A_2\| \leq \inf_{\Delta A \in S_{u1}} \|\Delta A_1 + \Delta A_2\| \\ &= \inf_{\Delta A_1 \in S_{u1}} \|\Delta A_1\| = \mu_1 \end{aligned}$$

and

$$\begin{aligned} \xi &= \inf_{\Delta A \in S_u} \|\Delta A_1 + \Delta A_2\| \leq \inf_{\Delta A \in S_{u2}} \|\Delta A_1 + \Delta A_2\| \\ &= \inf_{\Delta A_2 \in S_{u2}} \|\Delta A_2\| = \mu_2. \end{aligned}$$

Hence

$$\xi \leq \min(\mu_1, \mu_2). \quad (11)$$

On comparing (11) with (8), we see that ν for the F-M model is in a certain sense similar to ξ in the Roesser model. In the rest of the paper, our attention will be focused on investigating bound ν .

III. A LOWER BOUND OF ν AND ITS EVALUATION

A. A Lower Bound of ν

Let $\Omega = \{(\omega_1, \omega_2) : -\pi \leq \omega_1 \leq \pi, -\pi \leq \omega_2 \leq \pi\}$, and define

$$\gamma = \inf_{(\omega_1, \omega_2) \in \Omega} \underline{\sigma}[I_n - e^{j\omega_1} A_1 - e^{j\omega_2} A_2] \quad (12)$$

where I_n is the identity matrix of dimension n . As

$$\underline{\sigma}[I_n - e^{j\omega_1} A_1 - e^{j\omega_2} A_2] = \underline{\sigma}[e^{-j\omega_1} I_n - A_1 + e^{j(\omega_2 - \omega_1)} A_2]$$

where $\mu = \omega_2 - \omega_1$ and $\omega = -\omega_1$, we have

$$\begin{aligned} \gamma &= \inf_{-\pi \leq \mu \leq \pi} \inf_{-\pi \leq \omega \leq \pi} \underline{\sigma}[e^{j\omega} I_n - A(\mu)] \\ &= \inf_{0 \leq \mu \leq \pi} \inf_{-\pi \leq \omega \leq \pi} \underline{\sigma}[e^{j\omega} I_n - A(\mu)] \end{aligned} \quad (13)$$

where

$$A(\mu) = A_1 + e^{j\mu} A_2 \quad (14)$$

and the last equality in (13) is due to the fact that

$$f(\mu) = \inf_{-\pi \leq \omega \leq \pi} \underline{\sigma}[e^{j\omega} I_n - A(\mu)] \quad (15)$$

is an even function for $\mu \in [-\pi, \pi]$. With the definition of $f(\mu)$ in (15), (13) becomes

$$\gamma = \inf_{0 \leq \mu \leq \pi} f(\mu). \quad (16)$$

To explore the quantitative relation of γ to the robust stability measure ν , we note that (1) is asymptotically stable if and only if all eigenvalues of $A(\mu)$ for each $\mu \in [0, \pi]$ are strictly inside the unit circle [9]. Hence $A + \Delta A$ represents a stable system matrix if and only if $\Delta A \begin{bmatrix} I \\ e^{j\mu} I \end{bmatrix}$ is a stable perturbation of $A(\mu)$ for each $\mu \in [0, \pi]$. Since the exact 2-norm bound of complex, stable perturbations of $A(\mu)$ is equal to $f(\mu) = \min_{-\pi \leq \omega \leq \pi} \sigma[e^{j\omega} I - A(\mu)]$ [10]–[12], it follows that ΔA is a stable perturbation of A if

$$\left\| \Delta A \begin{bmatrix} I \\ e^{j\mu} I \end{bmatrix} \right\| < f(\mu) \text{ for all } \mu \in [0, \pi]. \quad (17)$$

As $\left\| \Delta A \begin{bmatrix} I \\ e^{j\mu} I \end{bmatrix} \right\| \leq \sqrt{2} \|\Delta A\|$ and $f(\mu) \geq \gamma$, (17) holds if

$$\|\Delta A\| < \frac{\gamma}{\sqrt{2}} \equiv \gamma_0. \quad (18)$$

This proves the following proposition.

Proposition 1: Define γ by (12). $\gamma_0 = \gamma/\sqrt{2}$ is a lower bound of the robust-stability measure ν defined by (6).

Two immediate consequences from above analysis are as follows.

Proposition 2: $\Delta A = [\Delta A_1 \quad \Delta A_2]$ is a stable perturbation for system (1) if

$$\|\Delta A_1\| + \|\Delta A_2\| < \gamma. \quad (19)$$

Proof: For any μ

$$\left\| \Delta A \begin{bmatrix} I \\ e^{j\mu} I \end{bmatrix} \right\| \leq \|\Delta A_1\| + \|\Delta A_2\|. \quad (20)$$

So if (19) holds, then $f(\mu) \geq \gamma$ leads (20) to (17). \square

Proposition 3: ΔA is a stable perturbation for system (1) if

$$\tilde{\gamma} < \gamma \quad (21)$$

where

$$\tilde{\gamma} = \sup_{0 \leq \mu \leq \pi} \|\Delta A_1 + e^{j\mu} \Delta A_2\|. \quad (22)$$

Proof: If (21) holds, then for any μ

$$\left\| \Delta A \begin{bmatrix} I \\ e^{j\mu} I \end{bmatrix} \right\| \leq \tilde{\gamma} < \gamma \leq f(\mu)$$

which leads to (17). \square

B. Computational Issues

We see from (13) that computing γ is a minimization problem with the objective function given by $F(\omega, \mu) = \sigma[e^{j\omega} I - A(\mu)]$. Because of the periodicity of $F(\omega, \mu)$ with respect to ω and μ , this minimization can be considered as an unconstrained problem, hence efficient optimization methods such as quasi-Newton algorithms can be used to find γ . As $F(\omega, \mu)$ represents the smallest singular value of a two-parameter matrix, there is in general no closed-form formula to compute the gradient of $F(\omega, \mu)$. For any $\delta > 0$, however, elementary properties of singular values [13, Chapter 6] imply that

$$|F(\omega + \delta, \mu) - F(\omega, \mu)| \leq |e^{j(\omega + \delta)} - e^{j\omega}| \leq \delta$$

and

$$|F(\omega, \mu + \delta) - F(\omega, \mu)| \leq \|A(\mu + \delta) - A(\mu)\| \leq \delta \|A_1\|$$

indicating that for any $\delta > 0$, ratio $[F(\omega + \delta, \mu) - F(\omega, \mu)]/\delta$ and $[F(\omega, \mu + \delta) - F(\omega, \mu)]/\delta$ are bounded by unity and $\|A_1\|$, respectively. Hence the gradient of $F(\omega, \mu)$ if it exists is bounded and reliable approximation of it can be obtained using simple numerical differentiations (i.e., the above difference ratio) with sufficiently small δ .

An alternative approach to the numerical evaluation of γ is to make use of (16) in conjunction with the fact that for a fixed μ , $f(\mu)$ defined by (15) is the exact bound for stable, complex perturbations of the system with $A(\mu)$ as its system matrix. Consequently, $f(\mu)$ for a fixed μ can be computed using the efficient bisection method proposed by Byers [14]. Again, due to the periodicity of $f(\mu)$, (15) is indeed a scalar minimization problem, which can be solved using, for example, the modified three-point-pattern algorithm [15, Section 7.3].

We shall touch upon these computational issues again in Section V through an illustrative example.

IV. LOWER BOUNDS DERIVED USING THE GENERALIZED 2-D LYAPUNOV EQUATION

A. The Generalized 2-D Lyapunov Equation

In [7] the following proposition was given.

Proposition 4 [7]: Equation (1) is asymptotically stable if there are positive definite P , W_1 , and W_2 such that

$$Q = \begin{bmatrix} P^{1/2} W_1 P^{1/2} & 0 \\ 0 & P^{1/2} W_2 P^{1/2} \end{bmatrix} - A^T P A > 0 \quad (23)$$

and that

$$I_n - W_1 - W_2 \geq 0. \quad (24)$$

This proposition is a generalization of the 2-D Lyapunov theorem proposed recently by Hinamoto [8]. It was shown in [7] that as compared to Hinamoto's result, the generalized Lyapunov equation (23) can be used to confirm stability of a broader class of 2-D discrete systems in the F-M state space setting. As [7] has not been available to the reader, for the sake of convenience a proof of Proposition 4 is given in Appendix A.

B. Two Lower Bounds of ν

Consider a stable system (1) that satisfies (23) and (24) with some positive definite P , W_1 and W_2 , and let the perturbed system be given by

$$x(i+1, j+1) = (A_1 + \Delta A_1)x(i, j+1) + (A_2 + \Delta A_2)x(i+1, j). \quad (25)$$

The matrices P , W_1 , and W_2 are employed to construct the Lyapunov function

$$\phi(i, j) = \phi_1(i, j) + \phi_2(i, j) \quad (26)$$

where

$$\phi_1(i, j) = x^T(i, j+1) P^{1/2} W_1 P^{1/2} x(i, j+1) \quad (27)$$

$$\phi_2(i, j) = x^T(i+1, j) P^{1/2} W_2 P^{1/2} x(i+1, j). \quad (28)$$

Define

$$\phi_{11}(i, j) = \phi_1(i+1, j) + \phi_2(i, j+1). \quad (29)$$

By (24)

$$\begin{aligned} \phi_{11}(i, j) &= x^T(i+1, j+1) P^{1/2} (W_1 + W_2) P^{1/2} x(i+1, j+1) \\ &\leq x^T(i+1, j+1) P x(i+1, j+1). \end{aligned} \quad (30)$$

Now using (23) and (25)–(30) we compute

$$\begin{aligned} \Delta \phi(i, j) &= \phi_{11}(i, j) - \phi(i, j) \\ &\leq \hat{x}^T(i, j) [(A^T + \Delta A) P (A + \Delta A) - \hat{P}] \hat{x}(i, j) \\ &\leq -[\sigma(Q) - 2\hat{\sigma}^{1/2}(\hat{P} - Q)\hat{\sigma}^{1/2}(P)] \|\Delta A\| \\ &\quad - \hat{\sigma}(P) \|\Delta A\|^2 \|\hat{x}(i, j)\|^2 \end{aligned}$$

where

$$\tilde{x}(i, j) = \begin{bmatrix} x(i, j+1) \\ x(i+1, j) \end{bmatrix} \quad \text{and} \quad \tilde{P} = \begin{bmatrix} P^{T/2} W_1 P^{1/2} & 0 \\ 0 & P^{1/2} W_2 P^{1/2} \end{bmatrix}. \quad (31)$$

Hence if

$$\|\Delta A\| < \frac{[\bar{\sigma}(\tilde{P} - Q) + \underline{\sigma}(Q)]^{1/2} - \bar{\sigma}^{1/2}(\tilde{P} - Q)}{\bar{\sigma}^{1/2}(P)} \quad (32)$$

then $\Delta\phi(i, j) < 0$ for all i, j , and a typical argument (see, e.g., [16]) applies to show the stability of the perturbed system (25).

An alternative Lyapunov function may be defined as

$$v(i, j) = [\phi_1(i, j) + \phi_2(i, j)]^{1/2}. \quad (33)$$

Similarly we define

$$v_{11}(i, j) = [\phi_1(i+1, j) + \phi_2(i, j+1)]^{1/2} \quad (34)$$

and use (24) to write

$$v_{11}(i, j) \leq [x^T(i+1, j+1) P x(i+1, j+1)]^{1/2}. \quad (35)$$

Now (23), (25), (33), and (35) imply that

$$\begin{aligned} \Delta v(i, j) &= v_{11}(i, j) - v(i, j) \\ &\leq [\tilde{x}^T(i, j)(A + \Delta A)^T P (A + \Delta A)\tilde{x}(i, j)]^{1/2} \\ &\quad - [\tilde{x}^T(i, j)\tilde{P}\tilde{x}(i, j)]^{1/2} \\ &\leq -\left[\frac{\underline{\sigma}(Q)}{\sigma^{1/2}(\tilde{P} - Q) + \sigma^{1/2}(P)} - \bar{\sigma}^{1/2}(P) \|\Delta A\| \right] \\ &\quad \|x(i, j)\|. \end{aligned}$$

Thus if

$$\|\Delta A\| < \frac{\underline{\sigma}(Q)}{\sigma^{1/2}(P)[\sigma^{1/2}(P - Q) + \bar{\sigma}^{1/2}(\tilde{P})]} \quad (36)$$

then $\Delta v(i, j) < 0$ for all i, j , and therefore, the perturbed system (25) will remain stable.

Although our numerical experience indicates that the lower bounds of ν given in (32) and (36) usually stay quite close each other, the bound given in (32) is always better than that in (36) as is shown in the next proposition.

Proposition 5.

$$\beta_{L,2} \leq \beta_{L,1} \quad (37)$$

where

$$\beta_{L,1} = \frac{[\bar{\sigma}(P - Q) + \underline{\sigma}(Q)]^{1/2} - \bar{\sigma}^{1/2}(P - Q)}{\bar{\sigma}^{1/2}(P)} \quad (38)$$

$$\beta_{L,2} = \frac{\underline{\sigma}(Q)}{\bar{\sigma}^{1/2}(P)[\bar{\sigma}^{1/2}(\tilde{P} - Q) + \sigma^{1/2}(\tilde{P})]} \quad (39)$$

with \tilde{P} given by (31).

The proof of the above proposition can be found in Appendix B.

C. Remarks on Numerical Solutions of the Generalized 2-D Lyapunov Equation

To compute lower bound $\beta_{L,1}$ or $\beta_{L,2}$, one must find positive definite P , W_1 , and W_2 that satisfy (23) and (24). Although constraints (24), $P > 0$, $W_1 > 0$ and $W_2 > 0$ can be put together as a set of linear matrix inequalities (LMI's), it is difficult to formulate the problem of solving (23) as a convex programming problem such as those investigated in [23] and [24], since solving (23) cannot be formulated as a problem of minimizing λ with $\lambda B(x) - A(x) > 0$ where x is the vector collecting all the parameters in P , W_1 , and W_2 , and $B(x)$ and $A(x)$ are matrices depending on x affinely. The approach taken here is to reformulate the problem as conventional

constrained minimization problem as follows. Note that (23) and (24) are equivalent to

$$\tilde{Q} \equiv I_{2n} - \begin{bmatrix} V_1^T & 0 \\ 0 & V_2^T \end{bmatrix} \tilde{A}^T \tilde{A} \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} > 0 \quad (40)$$

and

$$I - V_1^{-T} V_1^{-1} - V_2^{-T} V_2^{-1} \geq 0 \quad (41)$$

respectively, where $T = P^{-1/2}$, $\tilde{A} = [\tilde{A}_1 \ \tilde{A}_2]$, $\tilde{A}_i = T^{-1} A_i T$, $W_i = V_i^{-T} V_i^{-1}$ ($i = 1, 2$), which are in turn equivalent to

$$\underset{\substack{V_1, V_2 \\ \text{non-singular}}} \text{minimize} \quad \left\| \tilde{A} \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} \right\| < 1. \quad (42)$$

Solving the constrained minimization problem (42) is in general numerically intensive in the sense that it usually requires a fairly large amount of computations to obtain a local minimum, due primarily to the iterative nature of the algorithm. When the norm of system matrix A or the norm of a weighted version of A is small in a sense specified below, however, solutions to the generalized 2-D Lyapunov equation (23) and (24) can be found easily. For example, if

$$\|A\| < \frac{1}{\sqrt{2}} \quad (43)$$

then $W_1 = W_2 = \frac{1}{2} I_n$ and $P = I_n$ satisfy (24) and lead Q in (23) to

$$Q = \frac{1}{2} I_{2n} - A^T A$$

which is positive definite. The above choice of W_1 , W_2 , and P leads to $\tilde{P} = \frac{1}{2} I_{2n}$, $\bar{\sigma}(P - Q) = \|A\|^2$, $\underline{\sigma}(Q) = \frac{1}{2} - \|A\|^2$, and $\bar{\sigma}(P) = 1$. The two lower bounds in this case are found equal, being

$$\beta_{L,1} = \beta_{L,2} = \frac{1}{\sqrt{2}} - \|A\|. \quad (44)$$

Another simple case is when weights $\alpha > 0$ and $\beta > 0$ can be found such that

$$\frac{1}{\alpha^2} + \frac{1}{\beta^2} = 1 \quad (45)$$

and

$$\|[\alpha A_1 \ \beta A_2]\| < 1. \quad (46)$$

Obviously, $W_1 = \frac{1}{\alpha^2} I_n$, $W_2 = \frac{1}{\beta^2} I_n$, and $P = I_n$ satisfy (24) and lead Q in (23) to

$$Q = \begin{bmatrix} \frac{1}{\alpha^2} I & 0 \\ 0 & \frac{1}{\beta^2} I \end{bmatrix} - A^T A \quad (47)$$

which is positive definite if (46) holds. The two lower bounds in this case are given by

$$\beta_{L,1} = [\|A\|^2 + \underline{\sigma}(Q)]^{1/2} - \|A\| \quad (48)$$

and

$$\beta_{L,2} = \frac{\underline{\sigma}(Q)}{\|A\| + \max(\alpha^{-1}, \beta^{-1})}. \quad (49)$$

V. AN EXAMPLE

Consider a fourth-order F-M model with

$$A_1 = \begin{bmatrix} 0.0058 & 0.2853 & -0.2432 & 0.1246 \\ -0.0084 & 0.2377 & 0.1606 & -0.1404 \\ -0.0084 & -0.0217 & 0.1785 & 0.1662 \\ -0.0245 & 0.0884 & -0.0321 & 0.2428 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 0.0495 & 0.0687 & 0.2030 & -0.3013 \\ 0.1630 & 0.4817 & -0.0993 & 0.1814 \\ -0.2842 & 0.1790 & 0.5551 & 0.0799 \\ 0.2112 & -0.1887 & -0.0895 & 0.5604 \end{bmatrix}.$$

As $\|A\| = 0.7791$ exceeds $1/\sqrt{2} \approx 0.7071$, we attempt to find positive scalars α and β that satisfy (45) and (46). Using (45) we write $\beta = \alpha/\sqrt{\alpha^2 - 1}$ and define the objective function

$$e(\alpha) = \|[\alpha A_1 \quad \frac{\alpha}{\sqrt{\alpha^2 - 1}} A_2]\| \text{ for } 1 < \alpha < \infty. \quad (50)$$

Applying the modified three-point-pattern minimization algorithm [15, Chapter 7] to the bracket (i.e., an interval of α) [1.1, 3], it took 12 iterations with 17.796 Kflops (10^3 floating point operations) to find the minimum of $e(\alpha)$, $e(\alpha^*) = 0.9001 < 1$ at $\alpha^* = 2$. This leads to a solution to (23)–(24) with $W_1 = 0.25I_4$, $W_2 = 0.75I_4$, and $P = I_4$. It follows from Proposition 4 that the 2-D system is stable. From (47)–(49), lower bounds β_{L_1} and β_{L_2} were found to be $\beta_{L_1} = 0.0310$ and $\beta_{L_2} = 0.0300$. The total number of flops used in computing β_{L_1} or β_{L_2} is about 21 Kflops. To evaluate the lower bound β_0 in (18), a quasi-Newton optimization method was used to evaluate ν , where the gradient of $F(\omega, \mu)$ was performed numerically with $h = 10^{-6}$. It took nine iterations (with 241.6 Kflops) to obtain $\gamma = 0.1775$ at $[\omega^*, \mu^*] = [0.0252 \ 0.3856]$, which leads to $\beta_0 = 0.1255$. It is noted that bound β_0 is considerably less conservative as compared to β_{L_1} and β_{L_2} although computing β_0 is far more expensive in this case.

VI. CONCLUDING REMARKS

A robust stability measure has been proposed and three lower bounds of ν for F-M 2-D discrete systems have been derived by two distinct approaches. It appears from the numerical example discussed in Section V that the bounds β_{L_1} and β_{L_2} derived using the Lyapunov approach are quite conservative. This could be improved by carrying out the optimization problem (42) rather than using (50). As a matter of fact, the norm in (42) becomes $e(\alpha)$ in (50) if $T = I_1$, $V_1 = \alpha I_1$ and $V_2 = \frac{\alpha}{\sqrt{\alpha^2 - 1}} I_4$ with $\alpha > 1$ are chosen. With more parameters in (42), better lower bounds β_{L_i} ($i = 1, 2$) can be found at the expense of higher computation intensity.

APPENDIX A

Proof of Proposition 4: We prove the proposition by contradiction. Suppose the conditions of the proposition are satisfied but (1) is unstable. Then there is $(z_1, z_2) \in \bar{U}^*$ such that

$$\det(I_n - z_1 A_1 - z_2 A_2) = 0. \quad (51)$$

Hence there exists vector $v \neq 0$ such that

$$v = A \begin{bmatrix} z_1 I_n \\ z_2 I_n \end{bmatrix} v. \quad (52)$$

Using (23) and (52), we compute

$$v^* P v = v^* [\bar{z}_1 I_n \quad \bar{z}_2 I_n] \begin{bmatrix} P^{1/2} W_1 P^{1/2} & 0 \\ 0 & P^{1/2} W_2 P^{1/2} \end{bmatrix} \begin{bmatrix} z_1 I_n \\ z_2 I_n \end{bmatrix} v - v^* Q v,$$

where v^* denotes the complex conjugate transpose of v , \bar{z}_i is the complex conjugate of z_i ($i = 1, 2$), and $v_z^* = v^* [\bar{z}_1 I_n \quad \bar{z}_2 I_n]$. It follows that

$$v^* P^{1/2} (I_n - |z_1|^2 W_1 - |z_2|^2 W_2) P^{1/2} v = -v_z^* Q v. \quad (53)$$

By (51) we note that $(z_1, z_2) \neq (0, 0)$, hence $v_z \neq 0$. So $Q > 0$ means that the right-hand side of (53) is strictly negative. On the other side of equation, $|z_1| \leq 1$ and $|z_2| \leq 1$ and the positive semi-definiteness of $I_n - W_1 - W_2$ [see (24)] imply that $I - |z_1|^2 W_1 - |z_2|^2 W_2 \geq 0$. Therefore the left-hand side of (53) is nonnegative, leading to a contradiction. This completes the proof. \square

APPENDIX B

Proof of Proposition 5: Let $a = \underline{g}(Q)$, $b = \bar{\sigma}(P)$, $c = \bar{\sigma}(\bar{P} - Q)$, and $d = \bar{\sigma}(\bar{P})$. Note that $a > 0$, $b > 0$, $c \geq 0$, $d > 0$, and that

$$\bar{\sigma}(\bar{P} - Q) \leq \bar{\sigma}(\bar{P}) - \underline{g}(Q)$$

which implies that

$$(a + c)^{1/2} + c^{1/2} \leq c^{1/2} + d^{1/2}. \quad (54)$$

Since $b > 0$ and $d > 0$, it follows from (54) that

$$\frac{a}{b^{1/2}(c^{1/2} + d^{1/2})} \leq \frac{(a + c)^{1/2} - c^{1/2}}{b^{1/2}}. \quad \square$$

REFERENCES

- [1] E. Fornasini and G. Marchesini, "Doubly indexed dynamical systems: State-space models and structural properties," *Math. Syst. Theory*, vol. 12, pp. 59–72, 1978.
- [2] P. Agathoklis, "Estimation of the stability margin of 2-D discrete systems using the 2-D Lyapunov equation," in *Proc. IEEE Int. Symp. CAS*, Kyoto, Japan, vol. 3, 1985, pp. 1091–1092.
- [3] W.-S. Lu, A. Antoniou, and P. Agathoklis, "Stability of 2-D digital filters under parameter variations," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 476–482, May 1986.
- [4] P. Agathoklis, "Lower bounds for the stability margin of discrete two-dimensional systems based on the two-dimensional Lyapunov equation," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 745–749, June 1988.
- [5] W.-S. Lu, "Stability robustness of two-dimensional discrete systems and its computation," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 285–288, Feb. 1989.
- [6] —, "Some new results on stability robustness of 2-D discrete systems," *Multidimensional Syst. Sig. Process.* vol. 5, no. 4, pp. 345–361, 1994.
- [7] —, "On a Lyapunov approach to stability analysis of 2-D digital filters," *IEEE Trans. Circuits Syst. I*, vol. 41, pp. 665–669, Oct. 1994.
- [8] T. Hinamoto, "2-D Lyapunov equation and filter design based on Fornasini-Marchesini second model," *IEEE Trans. Circuits Syst. I*, vol. 40, pp. 102–110, Feb. 1993.
- [9] E. Fornasini and G. Marchesini, "Stability analysis of 2-D systems," *IEEE Trans. Circuits Syst.*, vol. 27, pp. 1210–1217, Dec. 1980.
- [10] W. H. Lee, *Robustness analysis for state space models*, Alpha-tech Inc., Rep. TP-151, 1982.
- [11] C. F. Van Loan, "How near is a stable matrix to an unstable matrix," *Contemporary Math.*, vol. 47, pp. 465–477, 1985.
- [12] J. M. Martin, "State space measure for stability robustness," *IEEE Trans. Automat. Contr.*, vol. 32, pp. 509–512, June 1987.
- [13] G. W. Stewart, *Introduction to Matrix Computations*, New York: Academic, 1973.
- [14] R. Byers, "A bisection method for measuring the distance of a stable matrix to the unstable matrices," *SIAM J. Sci. Statist. Comput.*, vol. 9, no. 5, pp. 875–881, 1988.
- [15] D. G. Luenberger, *Linear and Nonlinear Programming*, Reading, MA: Addison-Wesley, 1984.
- [16] N. G. El-Agizi and M. M. Fahmy, "Two-dimensional digital filters with no overflow oscillations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 465–469, 1979.
- [17] L. Qiu and E. J. Davison, "A new method for the stability robustness determination of state space models with real perturbations," in *Proc. IEEE Conf. Decis. Contr.*, Austin, TX, Dec. 1988, pp. 538–543.
- [18] D. Hinrichsen and A. J. Pritchard, "Real and complex stability radius: A survey," in *Control of Uncertain Systems*, D. Hinrichsen and B. Martensson, Eds., Boston, MA: Birkhäuser, 1990.
- [19] L. Qiu and E. J. Davison, "The stability robustness determination of state space models with real unstructured perturbations," *Math. Control Signals Syst.*, vol. 4, pp. 247–267, 1991.
- [20] —, "An improved bound on the real stability radius," in *Proc. Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 588–589.
- [21] I. Lewkowicz, "When are the complex and real stability radius equal," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 880–883, June 1992.
- [22] L. Qiu, B. Bernhardsson, A. Rantzer, E. J. Davison, P. M. Young, and J. C. Doyle, "On the real structured stability radius," in *Proc. 12th IFAC World Cong.*, vol. 8, Sydney, Australia, July 1993, pp. 71–78.
- [23] S. Boyd, V. Balakrishnan, E. Feron, and L. El Ghaoui, "Control system analysis and synthesis via linear matrix inequalities," in *Proc. Amer. Contr. Conf.*, San Francisco, CA, June 1993, pp. 2147–2154.
- [24] S. Boyd and L. El Ghaoui, "Method of centers for minimizing generalized eigenvalues," *Linear Algebra and Its Applications*, vol. 188, pp. 63–111, 1993.

Multivariable System Identification via Continued-Fraction Approximation

Rolf Johansson

Abstract—This paper presents theory for multivariable system identification using matrix fraction descriptions and the matrix continued fraction description approach which, in turn, yields a lattice-type order-recursive structure. Once the matrix continued-fraction expansion has been determined, it is straightforward to obtain solutions to both the left and right coprime factorizations of transfer function estimates and, in addition, a solution to problems of state estimation (observer design) and pole-assignment control. An important and attractive technical property is that calculation of transfer functions on the form of right and left coprime factorizations, calculation of state variable observers, and regulators all can be made using causal polynomial transfer functions defined by means of matrix sequences of the continued-fraction expansion applied in causal and stable forward-order and backward-order recursions.

1. INTRODUCTION

A long-standing and somewhat embarrassing problem in system identification is how to extend identification methods developed for single-input single-output systems to the case of multi-input multi-output systems. One important reason for this situation is that a large variety of ARMAX-type system parameterizations are suitable system descriptions and that no unique parameterization can be easily defined [13]–[14], [16], [18], [20]. This leaves a user with uncertainty as to the validity and numerical accuracy of calculations—e.g., state estimation and control, based on such transfer function estimates. For instance, consider the following two factorizations

$$\begin{cases} A_R(z^{-1})\xi = u & \text{and} \\ y = B_R(z^{-1})\xi \end{cases} \quad \begin{cases} A_L(z^{-1})Q(z^{-1})\xi = u \\ y = B_L(z^{-1})Q(z^{-1})\xi \end{cases} \quad Q \text{ unimodular} \quad (1)$$

where the partial state ξ provides a link to state-space realizations via the controllable canonical form [14, p. 403]. Clearly, the transfer functions of the two factorizations in (1) cannot be distinguished and there is no obvious choice of factorization to be preferred. As ξ is not known, there is also little hope to estimate (A_R, B_R) by ordinary methods of identification. As a result there are no general purpose methods nor software available whereas many control design procedure actually require coprime matrix fraction descriptions (MFD). As there are no established methods to find MFD's such as (1) from data, it remains a relevant problem in system identification to develop effective identification methods for such model structures [13], [16], [20]. A technical problem in this context is that statistical consistency of estimates of the matrix fractions becomes a meaningless issue as there is no unique parameter set towards which estimates may converge. Another as yet unsettled issue is the relationship between transfer function estimation, state estimation, and control system design.

A suitable starting point is to focus interest on transfer functions that are matrices of rational functions and transfer functions that can be approximated by a rational approximant (or a series of approximants for a sequence of increasing model orders). Approximation

Manuscript received March 20, 1994; revised July 6, 1994. This work was supported in part by the National Board for Industrial and Technical Development (NUTEK).

The author is with the Department of Automatic Control, Lund Institute of Technology, P.O. Box 118, S-22100, Lund, Sweden.

IEEE Log Number 9407222.

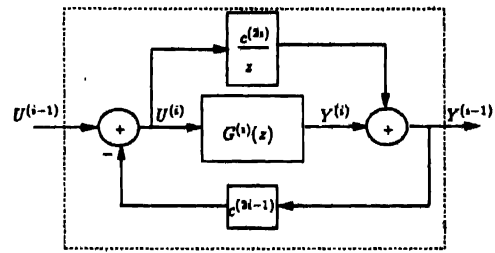


Fig. 1. Expansion of a transfer function $G^{(i-1)}(z)$ into the coefficients $c^{(2i-1)}$, $c^{(2i)}$ and a residual transfer function block $G^{(i)}(z)$ as practiced in continued fraction approximation.

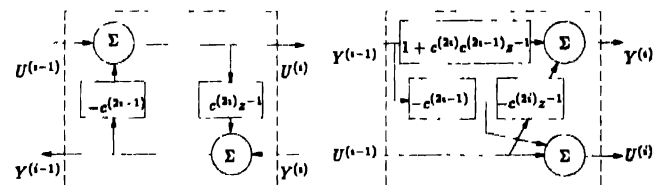


Fig. 2. Interpretation of the continued fraction expansion as a recursive transfer function relationship in a forward (right) or a forward/backward manner (left).

of complex-valued functions can be made by means of polynomials (Taylor series expansions) or rational functions (continued fraction approximation). Like infinite power series, continued fraction approximation can be used to represent certain types of analytic functions. Contrary to representations by power series, continued-fraction (CF) approximation may converge in regions that contain isolated singularities of the function to be represented and the approximants take on values in the extended complex plane $C \cup \{\infty\}$; see [8] and [23]. The method is well known from circuit theory and from the theory underlying the Routh criterion and the method has also been used for model reduction of linear systems [3]–[4], [6], [9]–[12], [15], multivariable systems [1]–[2], [7], and realization theory [2], [17], although recent focus in model reduction has been shifted towards methods based on balanced realization.

Consider an asymptotically stable system with a transfer function $G(z)$ in the forward shift operator or z -transform variable z . Then, the transfer function $\tilde{G}(z)$ may be expanded according to continued fraction which for a scalar $G(z)$ takes on the form

$$\begin{aligned} \tilde{G}(z) &= \frac{B(z^{-1})}{A(z^{-1})} = \frac{1}{c^{(1)} + \frac{1}{\frac{c^{(2)}}{z} + G^{(1)}(z)}} \\ &= \frac{1}{c^{(1)} + \frac{1}{\frac{c^{(2)}}{z} + \frac{1}{c^{(3)} + \frac{1}{\frac{c^{(4)}}{z} + G^{(2)}(z)}}}} \end{aligned} \quad (2)$$

with coefficients $\{c^{(j)}\}_{j=1}^{2m}$ —or more generally polynomials $\{c^{(j)}(z^{-1})\}_{j=1}^{2m}$ —and a residual transfer function block $G^{(m)}(z)$. The extension to multi-input multi-output systems is straightforward and even a block-diagram interpretation is possible; see Figs. 1–2. If $G^{(i)}(z) = 0$ (and thus $Y^{(i)} = 0$) at some stage i , the continued

fraction approximation terminates and results in a rational function approximant. Terminating continued fraction approximation are of particular importance in model reduction where methods can be proposed according to various principles of approximation—for instance, the approximation $G^{(m)}(z) \approx 0$, which truncates the sequence of coefficients after $2m$, $G^{(m)}(z) \approx G^{(m)}(1)$, which preserves the static gain of the system, or the Padé approximation which is known to sometimes providing poor approximants [12]. All such alternatives result in reduced-order models of order m with the approximant

$$H^{(m)}(z) = \frac{B^{(m)}(z^{-1})}{A^{(m)}(z^{-1})} \quad (3)$$

with continued-fraction coefficients obtained from polynomial coefficients by means of calculations reminiscent of those for the Routh criterion of stability. Recursions of the continued-fraction approximation therefore induce a ladder-type circuit theory reminiscent of lattice structures in spectral estimation (Fig. 2) and with the transfer matrix $M^{(i)}(z^{-1})$ relating stage i to stage $i-1$ in the order-recursive equation

$$\begin{pmatrix} Y^{(i)}(z) \\ U^{(i)}(z) \end{pmatrix} = M^{(i)}(z^{-1}) \begin{pmatrix} Y^{(i-1)}(z) \\ U^{(i-1)}(z) \end{pmatrix}$$

where

$$M^{(i)}(z^{-1}) = \begin{pmatrix} 1 + c^{(2i-1)}c^{(2i)}z^{-1} & -c^{(2i)}z^{-1} \\ -c^{(2i-1)} & 1 \end{pmatrix}. \quad (4)$$

A major difference between the present approach and the lattice filters used for whitening operations in signal processing is the property of its lattice transfer function

$$\begin{pmatrix} 1 & K_f z^{-1} \\ K_b & z^{-1} \end{pmatrix}, \quad K_f, K_b = \text{reflection coefficients} \quad (5)$$

which is not a unimodular matrix.

The main purpose of the present paper is to show how to obtain multivariable transfer function models from data on the form of coprime MFD's and to clarify the close relationship between the continued-fraction approximation and solutions to the Diophantine equation which is important in control system analysis for control design and observer design.

II. MULTIVARIABLE MODELING AND CONTROL

Let $\mathbb{R}[x]$ denote the polynomials in the indeterminate x with coefficients in \mathbb{R} , and let the set of $n \times m$ -matrices with entries in $\mathbb{R}[x]$ be denoted $\mathbb{R}^{n \times m}[x]$. Consider the input and output variables $U^{(i)} \in \mathbb{R}^m$, $Y^{(i)} \in \mathbb{R}^p$ for $i = 0, 1, \dots, n$ and matrices $C^{(2i)} \in \mathbb{R}^{p \times m}$, $C^{(2i-1)} \in \mathbb{R}^{m \times p}$ ($i = 1, \dots, n$) which exhibit the dependencies (Fig. 2)

$$\begin{aligned} Y^{(i-1)}(z) &= C^{(2i)}z^{-1}U^{(i)}(z) + Y^{(i)}(z) \\ U^{(i)}(z) &= -C^{(2i-1)}Y^{(i-1)}(z) + U^{(i-1)}(z) \end{aligned} \quad (6)$$

with a mixed backward/forward dependence in the recursion order i and with boundary values

$$\begin{pmatrix} Y^{(0)}(z) \\ U^{(0)}(z) \end{pmatrix} = \begin{pmatrix} Y(z) \\ U(z) \end{pmatrix} \quad (7)$$

and with z^{-1} denoting the backward shift operator or z -transform variable. A pure forward-order recursive equation for the continued fraction in (6) can be obtained as

$$\begin{pmatrix} Y^{(i)}(z) \\ U^{(i)}(z) \end{pmatrix} = \begin{pmatrix} I_{p \times p} + C^{(2i)}C^{(2i-1)}z^{-1} & -C^{(2i)}z^{-1} \\ -C^{(2i-1)} & I_{m \times m} \end{pmatrix} \cdot \begin{pmatrix} Y^{(i-1)}(z) \\ U^{(i-1)}(z) \end{pmatrix}. \quad (8)$$

Let $C^{(2i)}(x) \in \mathbb{R}^{p \times m}[x]$, $C^{(2i-1)}(x) \in \mathbb{R}^{m \times p}[x]$ for $i = 1, 2, \dots, n$. The polynomial matrix

$$M^{(i)}(x) = \begin{pmatrix} I_{p \times p} + C^{(2i)}C^{(2i-1)}x & -C^{(2i)}x \\ -C^{(2i-1)} & I_{m \times m} \end{pmatrix} \in \mathbb{R}^{(m+p) \times (m+p)}[x] \quad (9)$$

is unimodular with a unit determinant (i.e., $\det M^{(i)}(x) = 1$) and the inverse of $M^{(i)}(x)$ is

$$(M^{(i)}(x))^{-1} = \begin{pmatrix} I_{p \times p} & C^{(2i)}x \\ C^{(2i-1)} & I_{m \times m} + C^{(2i-1)}C^{(2i)}x \end{pmatrix}. \quad (10)$$

The matrix of (10) is unimodular since its inverse is also a polynomial matrix and from (10) is concluded that the order recursion of the continued fraction expansion is invertible so that

$$\begin{aligned} \begin{pmatrix} Y^{(i-1)}(z) \\ U^{(i-1)}(z) \end{pmatrix} &= (M^{(i)}(z^{-1}))^{-1} \begin{pmatrix} Y^{(i)}(z) \\ U^{(i)}(z) \end{pmatrix} \\ &= \begin{pmatrix} I_{p \times p} & C^{(2i)}z^{-1} \\ C^{(2i-1)} & I_{m \times m} + C^{(2i-1)}C^{(2i)}z^{-1} \end{pmatrix} \cdot \begin{pmatrix} Y^{(i)}(z) \\ U^{(i)}(z) \end{pmatrix}. \end{aligned} \quad (11)$$

An important technical property is that forward as well as backward order recursions are to be made using causal polynomial transfer functions which both are defined by means of the matrix sequences $\{C^{(2i-1)}\}$ and $\{C^{(2i)}\}$.

Now introduce the following notation for the matrix product

$$\begin{pmatrix} A_L^{(i)}(z^{-1}) & -B_L^{(i)}(z^{-1}) \\ S_L^{(i)}(z^{-1}) & R_L^{(i)}(z^{-1}) \end{pmatrix} = M^{(i)}(z^{-1})M^{(i-1)}(z^{-1}) \cdots M^{(1)}(z^{-1}) \quad (12)$$

and its inverse

$$\begin{pmatrix} R_R^{(i)}(z^{-1}) & B_R^{(i)}(z^{-1}) \\ -S_R^{(i)}(z^{-1}) & A_R^{(i)}(z^{-1}) \end{pmatrix} = (M^{(1)}(z^{-1}))^{-1} \cdots (M^{(i-1)}(z^{-1}))^{-1} (M^{(i)}(z^{-1}))^{-1} \quad (13)$$

with the backward polynomial transfer matrix $(M^{(i)}(z^{-1}))^{-1}$ defined by (10).

Lemma 1: Bezout identity and coprime factorizations

i) Any solution $X_L \in \mathbb{R}^{p \times p}[x]$ and $Y_L \in \mathbb{R}^{m \times p}[x]$ to the Diophantine equation

$$A_L^{(i)}(x)X_L(x) + B_L^{(i)}(x)Y_L(x) = I_{p \times p} \quad (14)$$

can be written

$$\begin{aligned} X_L &= R_R^{(i)} + B_R^{(i)}N \\ Y_L &= S_R^{(i)} - A_R^{(i)}N \end{aligned} \quad (15)$$

for some polynomial matrix $N \in \mathbb{R}^{m \times p}[x]$.

- ii) Any solution $X_R \in \mathbb{R}^{m \times m}[x]$ and $Y_R \in \mathbb{R}^{n \times p}[x]$ to the Diophantine equation

$$X_R(x)A_R^{(i)}(x) + Y_R(x)B_R^{(i)}(x) = I_{m \times m} \quad (16)$$

can be written

$$\begin{aligned} X_R &= R_L^{(i)} - N B_L^{(i)} \\ Y_R &= S_L^{(i)} + N A_L^{(i)} \end{aligned} \quad (17)$$

for some polynomial matrix $N \in \mathbb{R}^{m \times p}[x]$.

- iii) The matrix factorizations $\{(A_L^{(i)}(x), B_L^{(i)}(x))\}_{i=1}^n$ for $i = 1, 2, \dots, n$ are left coprime and the matrix factorizations $\{(A_R^{(i)}(x), B_R^{(i)}(x))\}_{i=1}^n$ for $i = 1, 2, \dots, n$ are right coprime.

A proof for the Bezout identity for general rings and—as a special case—for polynomial matrices is to be found in [22, Chapter 4]. According to the definitions in (12) and (13), it can be concluded that for any i we have the generalized Bezout identity (cf. [14, p. 382])

$$\begin{aligned} I_{(m+p) \times (m+p)} &= \begin{pmatrix} A_L^{(i)} & -B_L^{(i)} \\ S_L^{(i)} & R_L^{(i)} \end{pmatrix} \begin{pmatrix} R_R^{(i)} & B_R^{(i)} \\ -S_R^{(i)} & A_R^{(i)} \end{pmatrix} \\ &= \begin{pmatrix} A_L^{(i)} R_R^{(i)} + B_L^{(i)} S_R^{(i)} & A_L^{(i)} B_R^{(i)} - B_L^{(i)} A_R^{(i)} \\ S_L^{(i)} R_R^{(i)} - R_L^{(i)} S_R^{(i)} & S_L^{(i)} B_R^{(i)} + R_L^{(i)} A_R^{(i)} \end{pmatrix} \end{aligned}$$

with

$$\begin{aligned} \deg A_L^{(i)} &= \deg A_R^{(i)} = i \\ \deg B_L^{(i)} &= \deg B_R^{(i)} = i \\ \deg R_L^{(i)} &= \deg R_R^{(i)} = i - 1 \\ \deg S_L^{(i)} &= \deg S_R^{(i)} = i - 1 \\ A_L^{(i)}(0) &= R_R^{(i)}(0) = I_{p \times p} \\ A_R^{(i)}(0) &= R_L^{(i)}(0) = I_{m \times m} \end{aligned} \quad (18)$$

Note that particular and homogenous solutions to the Diophantine equation (14) can be extracted from (18) and that direct substitution of (15) into (14) gives

$$\begin{aligned} A_L^{(i)} X_L + B_L^{(i)} Y_L &= A_L^{(i)} (R_R^{(i)} + B_R^{(i)} N) + B_L^{(i)} (S_R^{(i)} - A_R^{(i)} V) \\ &= (A_L^{(i)} R_R^{(i)} + B_L^{(i)} S_R^{(i)}) + (A_L^{(i)} B_R^{(i)} + B_L^{(i)} A_R^{(i)}) N \\ &= I_{p \times p} + 0_{p \times m} \cdot N = I_{p \times p}. \end{aligned} \quad (19)$$

Coprime matrix fraction descriptions originating from terminating continued fraction expansions of polynomial order n are suitable for multivariable system modeling. Let the control object be represented as

$$\begin{aligned} \begin{pmatrix} A_L^{(n)}(z^{-1}) & -B_L^{(n)}(z^{-1}) \\ S_L^{(n)}(z^{-1}) & R_L^{(n)}(z^{-1}) \end{pmatrix} \begin{pmatrix} Y(z) \\ U(z) \end{pmatrix} &= \begin{pmatrix} Y^{(n)}(z) \\ U^{(n)}(z) \end{pmatrix} \\ \begin{pmatrix} R_R^{(n)}(z^{-1}) & B_R^{(n)}(z^{-1}) \\ -S_R^{(n)}(z^{-1}) & A_R^{(n)}(z^{-1}) \end{pmatrix} \begin{pmatrix} Y^{(n)}(z) \\ U^{(n)}(z) \end{pmatrix} &= \begin{pmatrix} Y(z) \\ U(z) \end{pmatrix} \end{aligned}$$

where

$$\begin{pmatrix} Y^{(n)}(z) \\ U^{(n)}(z) \end{pmatrix} = \begin{pmatrix} E(z) \\ \xi(z) \end{pmatrix} \quad (20)$$

with ξ as the partial state [14] and $E(z)$ as disturbance inputs and modeling error, $E(z)$ ideally being zero. In fact, (20) provides a means of correspondence between various system representations and, for instance, incorporates pole-assignment control which is easily calculated:

Theorem 1—Pole-Assignment to the Origin: Assume that the control object be described by (20). The control laws

$$\begin{aligned} i) \quad U(z) &= -R_L^{-1}(z^{-1}) S_L(z^{-1}) Y(z) \\ &\Rightarrow (S_L \quad R_L) \begin{pmatrix} Y \\ U \end{pmatrix} = 0_{n \times 1} \\ ii) \quad \begin{cases} R_R(z^{-1}) \xi_R(z) = Y(z) \\ U(z) = -S_R(z^{-1}) \xi_R(z) \end{cases} \\ &\Rightarrow \begin{pmatrix} R_R \\ -S_R \end{pmatrix} \xi_R = \begin{pmatrix} Y \\ U \end{pmatrix} \end{aligned} \quad (21)$$

both achieve pole assignment to the origin and $\xi = 0$.

Proof: For control law i) is found that

$$\begin{aligned} 0_{m \times 1} &= (S_L \quad R_L) \begin{pmatrix} Y \\ U \end{pmatrix} \\ &= (S_L \quad R_L) \cdot \begin{pmatrix} R_R & B_R \\ -S_R & A_R \end{pmatrix} \begin{pmatrix} E \\ \xi \end{pmatrix} \\ &= (S_L R_R - R_L S_R \quad S_L B_R + R_L A_R) \begin{pmatrix} E \\ \xi \end{pmatrix} \\ &= (0 \quad I_{m \times m}) \begin{pmatrix} E \\ \xi \end{pmatrix} = \xi \end{aligned} \quad (22)$$

with pole polynomial $S_L B_R + R_L A_R = I_{m \times m}$ as follows from (18) and the Bezout identity. Similarly, for control law ii) is found that with system dynamics determined from

$$\begin{aligned} \begin{pmatrix} E \\ \xi \end{pmatrix} &= \begin{pmatrix} A_L & -B_L \\ S_L & R_L \end{pmatrix} \begin{pmatrix} Y \\ U \end{pmatrix} \\ &= \begin{pmatrix} A_L & -B_L \\ S_L & R_L \end{pmatrix} \begin{pmatrix} R_R \\ -S_R \end{pmatrix} \xi_R \\ &= \begin{pmatrix} A_L R_R + B_L S_R \\ S_L R_R - R_L S_R \end{pmatrix} \xi_R \\ &= \begin{pmatrix} I_{p \times p} \\ 0_{m \times p} \end{pmatrix} \xi_R \end{aligned} \quad (23)$$

with the pole polynomial $A_L R_R + B_L S_R = I_{p \times p}$. Rearrangement gives

$$\begin{pmatrix} \xi_R \\ \xi \end{pmatrix} = \begin{pmatrix} E \\ 0_{m \times 1} \end{pmatrix} \quad (24)$$

which effectively decouples the state ξ from perturbation E . Moreover, the partial state ξ_R of the control law ii) provides an observer for the disturbance E . \square

It can be concluded that matrices $\{C^{(2i-1)}\}$, $\{C^{(2i)}\}$ with real-valued elements generate left and right coprime factorizations of a strictly proper control object and associated controllers. Using system equivalence results [21, pp. 73-74] there exist procedures to generate the corresponding controllable and observable state-space realizations and controllability indices and observability indices can be obtained [21, p. 51].

III. A MULTIVARIABLE IDENTIFICATION ALGORITHM

It remains to determine the continued fraction expansion matrices $\{C^{(2i-1)}\}_{i=1}^n$, $\{C^{(2i)}\}_{i=1}^n$ for nonunique transfer function estimates originating from standard identification methods—e.g., least-squares

identification For the case of least-squares identification, it is suitable to organize model and data sequences $\{y_k\}$ and $\{u_k\}$ according to

$$\begin{aligned} y_k &= -A_1 y_{k-1} - \dots - A_n y_{k-n} + B_1 u_{k-1} + \dots + B_n u_{k-n}, \\ y_k &\in \mathbb{R}^p \\ \phi_k &= (-y_{k-1}^T \dots -y_{k-n}^T \quad u_{k-1}^T \dots u_{k-n}^T)^T, \\ \phi_k &\in \mathbb{R}^{n(m+p)} \\ \theta &= (A_1 \quad \dots \quad A_n \quad B_1 \quad \dots \quad B_n)^T \\ \theta &\in \mathbb{R}^{n(m+p) \times p} \end{aligned} \quad (25)$$

which for N samples and model order n suggests the linear regression model

$$\mathcal{M} \quad \mathcal{Y}_N = \Phi_N \theta$$

with

$$\mathcal{Y}_N = \begin{pmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{pmatrix}$$

and

$$\Phi_N = \begin{pmatrix} \phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_N^T \end{pmatrix} \quad (26)$$

As a result of the nonuniqueness of parameters the normal equations of the associated least-squares estimation of θ may exhibit rank deficit and it is natural to apply the least squares solution

$$\theta_N = (\Phi_N^T \Phi_N)^+ \Phi_N^T \mathcal{Y}_N \quad (27)$$

where $(\Phi_N^T \Phi_N)^+$ denotes the matrix pseudo inverse of $\Phi_N^T \Phi_N$. The associated least-squares estimate thus obtained has the smallest 2-norm of all possible minimizers of the least-squares criterion and provides its result in the form of a left matrix factorization

$$\begin{aligned} A(z^{-1})Y(z) &= B(z^{-1})U(z) + W(z) \\ A(z^{-1}) &= I_{p \times p} + A_1 z^{-1} + \dots + A_n z^{-n} \\ B(z^{-1}) &= B_1 z^{-1} + B_2 z^{-2} + \dots + B_n z^{-n} \end{aligned}$$

with input U , output Y , and disturbance W .

Algorithm 1 Let $A_i^{(n)}(x) = I_{p \times p} + \sum_{j=1}^n A_j^{(n)} x^j \in \mathbb{R}^{p \times p}[x]$ and $B_L^{(n)}(x) = \sum_{j=1}^n B_j^{(n)} x^j \in \mathbb{R}^{p \times m}[x]$ denote some nonsingular left MFD $(A(x), B(x))$ of polynomial degree n . Compute for $i = n, n-1, \dots, 1$

$$\begin{aligned} i) \quad & C_j^{(2(n-i+1)-1)} \\ &= (B_{i-j}^{(i)})^+ \left(A_i^{(i)} - \sum_{k=0}^{j-1} B_{i-k}^{(i)} C_k^{(2(n-i+1)-1)} \right), \\ & \quad j = 0, 1, \dots, i-1 \end{aligned}$$

$$\begin{aligned} ii) \quad & C_j^{(2(n-i+1))} \\ &= (A_{i-j-1}^{(i-1)})^+ \left(B_i^{(i)} - \sum_{k=0}^{j-1} A_{i-1-k}^{(i-1)} C_k^{(2(n-i+1))} \right), \\ & \quad j = 0, 1, \dots, i-1 \end{aligned}$$

$$\begin{aligned} iii) \quad & C_j^{(2(n-i+1))}(x) \\ &= \sum_{j=0}^i C_j^{(2(n-i+1))} x^j \in \mathbb{R}^{m \times p}[x], \\ & \quad i = n, n-1, \dots, 1 \end{aligned}$$

$$\begin{aligned} iv) \quad & C_j^{(2(n-i+1)-1)}(x) \\ &= \sum_{j=0}^i C_j^{(2(n-i+1)-1)} x^j \in \mathbb{R}^{p \times p}[x] \\ & \quad i = n, n-1, \dots, 1 \end{aligned}$$

$$\begin{aligned} v) \quad & A_i^{(i-1)}(x) \\ &= A_i^{(i)}(x) - B_i^{(i)}(x) C^{(2(n-i+1)-1)}(x), \\ & \quad A_i^{(n)} = A(x) \end{aligned}$$

$$\begin{aligned} vi) \quad & B_i^{(i-1)}(x) \\ &= B_i^{(i)}(x) - A_i^{(i-1)}(x) C^{(2(n-i+1))}(x) x \\ & \quad B_i^{(n)} = B(x) \end{aligned} \quad (28)$$

where $(\cdot)^+$ denotes a generalized matrix inverse \square

Theorem 2—Termination of Continued Fraction Expansion Let $(A(x), B(x))$ be a strictly proper nonsingular left MFD of polynomial order n with $A(0) = I_{p \times p}$. Algorithm 1 determines the terminating continued fraction expansion matrices $\{C^{(i-1)}(x)\}_{i=1}^n$, $\{C^{(2i)}(x)\}_{i=1}^n$ and unimodular matrices $\{M^{(i)}(x)\}_{i=1}^n$ of Lemma 1 in no more than n steps with

$$\begin{aligned} Q(x)(I_{p \times p} - 0_{p \times n}) &= (A_L^{(0)} - B_L^{(0)}) = (A(x) - B(x)) \\ &= (M^{(1)}(x))^{-1} \dots (M^{(n)}(x))^{-1} (M^{(n)}(x))^{-1} \end{aligned} \quad (29)$$

with $Q(x) \in \mathbb{R}^{p \times p}[x]$ being a left divisor of $(A(x), B(x))$.

Proof Steps v)–vi) of the algorithm embody for $i = n, n-1, \dots, 1$ the order recursion

$$\begin{aligned} & (A_i^{(i-1)} - B_L^{(i-1)}) \\ &= (A_i^{(i)} - B_i^{(i)}) \\ & \quad \left(C^{(2(n-i+1)-1)} I + C^{(2(n-i+1)-1)} C^{(2(n-i+1))} C^{(2(n-i+1)-1)} \right) \end{aligned} \quad (30)$$

The construction similar to a Routh scheme of $\{C^{(2i-1)}(x)\}_{i=1}^n$, $\{C^{(2i)}(x)\}_{i=1}^n$ and the full-rank condition of $(A(x), B(x))$ serve to eliminate terms of degree i from $A_i^{(i-1)}$ and $B_i^{(i-1)}$ so that

$$\begin{aligned} 0 &= A_i^{(i)} x^i - \sum_{k=0}^{i-1} B_{i-k}^{(i)} C_k^{(2(n-i+1)-1)} x^i \\ 0 &= B_i^{(i)} x^i - \sum_{k=0}^{i-1} A_{i-1-k}^{(i-1)} C_k^{(2(n-i+1))} x^i \end{aligned} \quad (31)$$

as long as $B_i^{(i)} \neq 0$. The order recursion of $B_i^{(i)}$ for $i = n, n-1, \dots, 1$ with $\deg B_i^{(i)}(x) = i$ and $B_i^{(0)} = 0_{p \times n}$ and can always be made as $A_L^{(i)}(0) = I_{p \times p}$ for all i . The resultant $A_i^{(0)}$ terminates as a nonsingular left divisor of $(A(x), B(x))$ as $\{M^{(i)}(x)\}_{i=1}^n$ of (29) are unimodular \square

After determination of $C^{(2i-1)}$ and $C^{(2i)}$ in (30) for $i = n, n-1, \dots, 1$, it is straightforward to remove the left divisor $Q(x)$ and to recursively calculate the approximants $\{(A_i^{(i)}, B_i^{(i)})\}_{i=1}^n$ of polynomial degree i according to (12)–(13). The resultant approximant

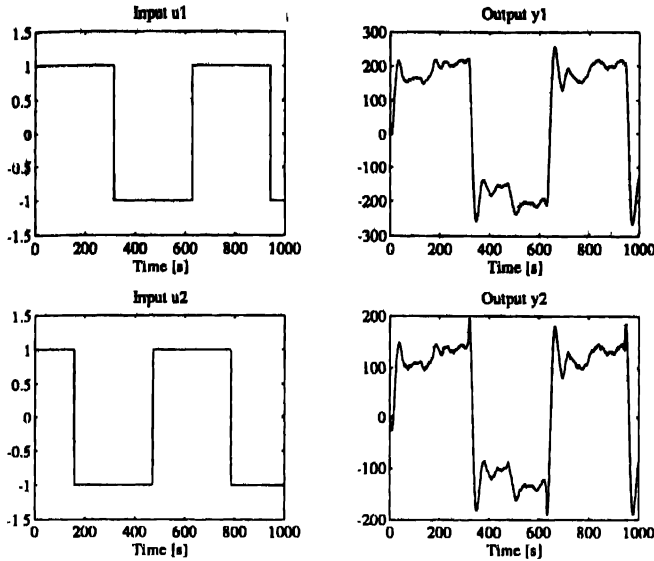


Fig. 3. Input-output data used in Example 2 with apparent cross-coupling properties.

represent coprime factorizations according to Theorem 1 with exact representation of the original transfer function for $i = n$. In particular, the sequences $\{C^{(2i)}\}$, $\{C^{(2i-1)}\}$ are constant matrices whenever the matrix sequences $\{B_i^{(1)}\}_{i=1}^n$ and $\{A_i^{(1)}\}_{i=1}^n$ are full-rank matrices.

Example 1—Common Factors: Consider the following left matrix factorization with the common left divisor $Q(z^{-1}) = I + Dz^{-1}$ (D being assumed invertible)

$$\begin{aligned} A_L^{(2)} &= (I + Dz^{-1})(1 + A_1 z^{-1}) \\ &= I + (D + A_1)z^{-1} + DA_1 z^{-2} \\ B_L^{(2)} &= (I + Dz^{-1})B_1 z^{-1} \\ &= B_1 z^{-1} + DB_1 z^{-2}. \end{aligned} \quad (32)$$

Application of the algorithm gives

$$\begin{aligned} C^{(1)} &= (DB_1)^+ DA_1 \\ A_L^{(1)} &= I + (D + A_1)z^{-1} + DA_1 z^{-2} \\ &\quad - (B_1 z^{-1} + DB_1 z^{-2})C^{(1)} = I + Dz^{-1} \\ C^{(2)} &= D^{-1}DB_1 = B_1 \\ B_L^{(1)} &= B_1 z^{-1} + DB_1 z^{-2} \\ &\quad - (I + Dz^{-1})C^{(2)} z^{-1} = 0. \end{aligned} \quad (33)$$

As $B_L^{(1)} = 0$ this algorithm terminates after one recursion only with the resulting model order $n - 1 = 1$ of the transfer function approximant. The common left divisor $(I + Dz^{-1})$ appears as the approximant $A_L^{(1)}$ as shown in Theorem 3. \square

Example 2—A Simulated Example: Consider the multivariable system

$$y_k = \begin{pmatrix} 1.2 & -0.29 \\ 0.25 & 0.7 \end{pmatrix} y_{k-1} + \begin{pmatrix} -2.0 & 0.07 \\ -10.0 & 1.0 \end{pmatrix} u_{k-1} + w_k, \quad u_k, y_k, w_k \in \mathbb{R}^2 \quad (34)$$

with input $\{u_k\}$, zero-mean white noise $\{w_k\}$ with $\mathcal{E}\{w_i w_j^T\} = I_{2 \times 2} \delta_{ij}$ and output $\{y_k\}$. Simulated data are according to Fig. 3, which exhibit obvious cross-coupling properties. The estimate $\hat{\theta}_N$ for model order $n = 1$ and $N = 1000$ samples of input-output data is

$$\begin{aligned} \hat{\theta}_N &= (\hat{A}_1 \quad \hat{B}_1)^T \\ &= \begin{pmatrix} -1.2012 & 0.2908 & -2.1053 & 0.0916 \\ -0.2490 & -0.7010 & -9.9715 & 1.0123 \end{pmatrix}^T \end{aligned} \quad (35)$$

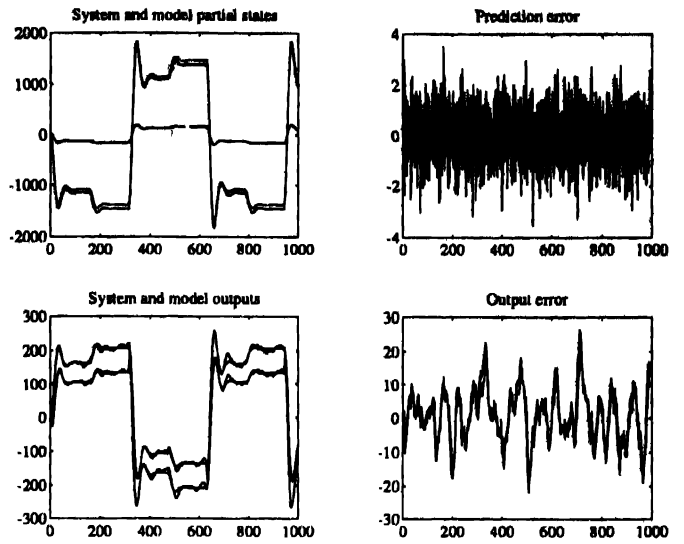


Fig. 4. Simulated data including partial state, output error, prediction error, and model output with input-output data as used in Example 2.

and the continued-fraction approximation matrices are

$$C^{(1)} = \begin{pmatrix} 0.9797 & -0.2945 \\ 9.4039 & -3.5929 \end{pmatrix}$$

and

$$C^{(2)} = \begin{pmatrix} -2.1053 & 0.0916 \\ -9.9715 & 1.0123 \end{pmatrix}. \quad (36)$$

Polynomial matrices of the right coprime factorization can be computed from the real-valued matrices $C^{(1)}$ and $C^{(2)}$ as

$$\begin{aligned} \begin{pmatrix} B_R(z^{-1}) \\ A_R(z^{-1}) \end{pmatrix} &= \begin{pmatrix} C^{(2)} z^{-1} \\ I + C^{(1)} C^{(2)} z^{-1} \end{pmatrix} \\ &= \begin{pmatrix} -2.1053 z^{-1} & 0.0916 z^{-1} \\ -9.9715 z^{-1} & 1.0123 z^{-1} \\ 1.0 + 0.8736 z^{-1} & -0.2084 z^{-1} \\ 16.0285 z^{-1} & 1.0 - 2.7759 z^{-1} \end{pmatrix}. \end{aligned} \quad (37)$$

Once the continued-fraction matrices $\{C^{(2i-1)}\}_{i=1}^n$ and $\{C^{(2i)}\}_{i=1}^n$ have been determined it is straightforward to obtain the prediction error ε and the partial state ξ according to

$$\begin{aligned} \begin{pmatrix} \varepsilon \\ \xi \end{pmatrix} &= \begin{pmatrix} \hat{A}_L^{(n)} & -\hat{B}_L^{(n)} \\ \hat{S}_L^{(n)} & \hat{R}_L^{(n)} \end{pmatrix} \begin{pmatrix} Y(z) \\ U(z) \end{pmatrix} \\ &= \hat{A}^{(n)} \hat{A}^{(n-1)} \dots \hat{A}^{(1)} \begin{pmatrix} Y(z) \\ U(z) \end{pmatrix}. \end{aligned} \quad (38)$$

The partial state ξ and up to n shifted values of ξ determine the state of the system described according to the controllable canonical state-space realization [14, p. 403], [2]. Thus, the state observer embodied in (38) provides a state estimate that is obtained along with the residual (or prediction error) of transfer function estimation. Simulated data including partial state, output error, prediction error, and model output are shown in Fig. 4.

IV. CONCLUSIONS

We have shown that the continued-fraction expansion provides an alternative parameterization of multivariable systems in the context of system identification. The close relationship between the continued-fraction approximation and solution to the Diophantine equation is

demonstrated. Hence, this parameterization provides a computational link between various matrix factorizations, between transfer function estimation and state estimation, between prediction-error methods and output-error methods, and between model reduction and control. The lattice structure provided by the continued-fraction approximation facilitates calculations and computations can be organized by means of order recursions backwards and forwards. In future work it is intended to investigate identification properties in the presence of colored noise and adaptive control.

REFERENCES

- [1] B. D. O. Anderson, "An approach to multivariable system identification," *Automatica*, vol. 13, pp. 401–408, 1977.
- [2] A. C. Antoulas and R. H. Bishop, "Continued-fraction decomposition of linear systems in the state space," *Syst. Contr. Lett.*, vol. 9, pp. 43–53, 1987.
- [3] C. F. Chen, "Model reduction of multivariable control systems by means of matrix continued fractions," *Int. J. Contr.*, vol. 20, pp. 225–238, 1974.
- [4] C. F. Chen and L. S. Shieh, "A novel approach to model simplification," *Int. J. Contr.*, vol. 8, pp. 561–570, 1968.
- [5] P. A. Fuhrmann, "A matrix Euclidean algorithm and matrix continued fraction expansions," *Syst. Contr. Lett.*, vol. 3, pp. 263–271, 1983.
- [6] M. J. Goldman, W. J. Porras, and C. T. Leondes, "Multivariable system reduction via Cauchy forms," *Int. J. Contr.*, vol. AC-34, pp. 623–650, 1981.
- [7] R. Guidorzi, "Canonical structures in the identification of multivariable systems," *Automatica*, vol. 11, pp. 361–374, 1975.
- [8] P. Henrici, *Applied and Computational Complex Analysis*, vol. 2. New York: Wiley, 1977.
- [9] C. Hwang and M.-Y. Chen, "Stable linear system reduction via a multipoint squared-magnitude continued-fraction expansion," in *IEE Proc. Part D: Control Theory and Applications*, vol. 135, no. 6, Nov. 1988, pp. 441–444.
- [10] C. Hwang and Y. C. Lee, "Multifrequency Padé approximation via Jordan continued fraction expansion," *IEEE Trans. Automat. Contr.*, vol. AC-34, pp. 444–446, 1989.
- [11] C. Hwang and M. Y. Chen, "A multipoint continued fraction expansion for linear system reduction," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 648–651, 1986.
- [12] —, "Solution of general Padé fitting problem via continued-fraction expansion," *IEEE Trans. Automat. Contr.*, vol. AC-32, no. 1, pp. 57–59, Jan. 1987.
- [13] R. Johansson, *System Modeling and Identification*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
- [14] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1980.
- [15] A. Linnemann, "Class of single-input single-output systems stabilizable by reduced-order controllers," *Syst. Contr. Lett.*, vol. 11, no. 1, pp. 27–32, July 1988.
- [16] L. Jung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
- [17] S. H. Mentzelopoulou and N. J. Theodorou, "n-dimensional minimal state-space realization," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 340–343, Mar. 1991.
- [18] L. Perneho, "An algebraic theory for the design of controllers for multivariable systems—Part I: Structure matrices and feedforward design; Part II: Feedback realizations and feedback design," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 171–182, 183–194, 1981.
- [19] L. S. Shieh, J. M. Navarro, and R. E. Yates, "Multivariable systems identification via frequency responses," *Proc. IEEE*, vol. 62, pp. 1169–1171, 1974.
- [20] T. Söderström and P. Stoica, *System Identification*. London: Prentice-Hall Int., 1989.
- [21] A. I. Vardulakis, *Linear Multivariable Control—Algebraic Analysis and Synthesis Methods*. Chichester, UK: Wiley, 1991.
- [22] M. Vidyasagar, *Control System Synthesis—A Factorization Approach*. Cambridge, MA: The MIT Press, 1985.
- [23] H. S. Wall, *Analytic Theory of Continued Fractions*. New York: Van Nostrand, 1948.

All Fixed-Order H_∞ Controllers: Observer-Based Structure and Covariance Bounds

T. Iwasaki and R. E. Skelton

Abstract—This note obtains a parameterization of the set of all stabilizing controllers of order less than or equal to the plant, which yields for the closed-loop transfer matrix a specified H_∞ norm bound. The algebraic results of covariance control are applied to the H_∞ control problem to yield a parameterization in terms of the Lyapunov matrix, which carries many system properties (such as H_2 performance, covariance bounds, system entropy at infinity, etc.). All low-order H_∞ controllers are shown to have observer-based structure for "reduced-order models" of the plant and are characterized by two Riccati equations with a coupling condition.

I. INTRODUCTION

The main objective of this note is to provide a new derivation of the standard H_∞ control theory [1], [2] using the algebraic approach in the covariance control literature [3]–[6]. The idea is to consider the bounded real equation for the closed-loop system as an algebraic problem to be solved for the controller parameters. Similar algebraic approaches in the literature include [4], [7]–[10]. In particular, [7] and [4] (appeared simultaneously) extend the result of [9] to obtain a parameterization of fixed-order output feedback H_∞ controllers. The "standard" case (full control penalty and full measurement noise) was considered in [7] where the existence conditions are characterized by coupled Riccati equations similar to those of [9]. These Riccati equations contain controller parameters, however, and thus the parameterization is implicit. Reference [4] (see [5] for its journal version) solved a singular case (no control penalty and no measurement noise) where existence conditions are given by three matrix equalities without controller parameters, and all fixed-order H_∞ controllers are parameterized explicitly. In [7] and [4], computations of fixed-order controllers are extremely difficult.

Contributions of this note over the existing literature are the following. We consider the "standard case" and remove the controller parameter dependence of the Riccati equations in [7] and obtain explicit formulas for all H_∞ controllers, which allow us to prove the observer-based structure of all H_∞ controllers, including noncentral controllers of order equal to or less than the plant (note that [1] shows the observer-based structure for the central (full-order) controller, but not for any other strictly proper H_∞ controllers). Another contribution is extending the results of [7], [4], to provide a parameterization of all H_∞ controllers in terms of the solution to the bounded real equation, which we call the H_∞ Lyapunov matrix. Recall that the H_∞ Lyapunov matrix is an upper bound for the state covariance [11] and can be used to define several other system properties including system entropy [12] and the (dual) mixed H_2/H_∞ performance measure defined using power semi-norms [13]. (An improved upper bound on H_2 norm is provided in the mixed H_2/H_∞ solutions in [14].) Thus, such a parameterization is useful (especially for the full-order controller case) to incorporate H_2 related performance specifications in addition to the H_∞ norm bound.

Manuscript received September 14, 1992; revised May 18, 1994.

The authors are with the Space Systems Control Laboratory Potter Engineering Center, Purdue University, West Lafayette, IN 47907-1293 USA.
IEEE Log Number 9407232.

Finally, we should mention that our result still has a computational difficulty, as in [7] and [4], for solving the coupled Riccati equations except for the full-order controller case. Thus the computational aspects of the fixed (low) order H_∞ controller design problem needs a more thorough exploration.

We will use the following notation. For a matrix A , A' denotes the transpose, $\rho(A)$ and $\sigma_{\max}(A)$ the spectral radius and the maximum singular value, respectively, and A^+ denotes the Moore-Penrose pseudo-inverse of A . If A is nonnegative definite, $A^{1/2}$ is the (unique) nonnegative definite square root of A . $\|\cdot\|_\infty$ and $\|\cdot\|_2$ denote the H_∞ and H_2 norms, respectively.

II. PRELIMINARIES

Consider the class of dynamic systems described by the state space realizations of the n_p -th-order plant (Σ_p) and the n_c -th-order controller (Σ_c)

$$(\Sigma_p) \begin{cases} \dot{x} = A_1 x + B_1 w + B_2 u \\ z = C_1 x + D_{12} u \\ y = C_2 x + D_{21} w \end{cases}$$

$$(\Sigma_c) \begin{cases} \dot{x}_c = A_c x_c + B_c y \\ u = C_c x_c + D_c y \end{cases}$$

to give the closed-loop system (Σ_{cl})

$$(\Sigma_{cl}) \begin{cases} \dot{x}_{cl} = A_{cl} x_{cl} + B_{cl} w \\ z = C_{cl} x_{cl} + D_{cl} w \end{cases} \quad x_{cl} := \begin{bmatrix} x \\ x_c \end{bmatrix}$$

where $x \in \mathbb{R}^{n_p}$ is the plant state, $w \in \mathbb{R}^{n_w}$ is any external input, including plant disturbances, measurement noise, etc., $u \in \mathbb{R}^{n_u}$ is the control input, $z \in \mathbb{R}^{n_z}$ is the regulated output, $y \in \mathbb{R}^{n_y}$ is the measured output, and $x_c \in \mathbb{R}^{n_c}$ is the controller state. We denote the closed-loop transfer matrix from w to z by T_{zw} . The following assumptions are made for (Σ_p) :

AP1) (A_1, B_1) is stabilizable and (C_2, A_1) is detectable.

AP2) (A_1, B_1) is stabilizable.

AP3) $D_{12}' [C_1' D_{12}] = [0 \ I']$, $I' > 0$
 $D_{21} [B_1' D_{21}'] = [0 \ V']$, $V' > 0$.

AP1) is obviously necessary for the existence of a stabilizing controller. AP2) and AP3) are technical assumptions which are standard in the H_∞ control literature (e.g., [1]). Note that AP2) and AP3) imply that (A_1, B_1) is stabilizable if and only if (A_1, B_1) is stabilizable. This fact can easily be verified by standard manipulations once we notice that

$$[B_1 \ 0] = \begin{bmatrix} B_1 & B_2 D_c \\ 0 & B_c \end{bmatrix} T, \quad T := \begin{bmatrix} I & D_{21}' \\ D_{21} & 0 \end{bmatrix}$$

where T is nonsingular due to the second part of AP3). To state the H_∞ control problem, we need the following definition.

Definition 1: Given a positive scalar γ , the controller (Σ_c) is said to be an H_∞ controller if the following four conditions hold:

C1) (A_c, B_c) is controllable and (A_c, C_c) is observable.

C2) A_c is asymptotically stable.

C3) $\|T_{zw}\|_\infty \leq \gamma$.

C4) $\gamma^2 I - D_{cl}' D_{cl} > 0$.

The first minimality condition reflects the fact that we seek a low-order controller. Note that conditions C2) and C3) imply $\gamma^2 I - D_{cl}' D_{cl} \geq 0$. To utilize Lemma 1 below, we impose condition C4) which excludes the case $\sigma_{\max}(D_{cl}) = \gamma$. Nevertheless, the set of

all stabilizing minimal controllers which yield the strict inequality $\|T_{zw}\|_\infty < \gamma$ is contained in the set of all H_∞ controllers defined above. Without loss of generality, we will set the H_∞ norm bound γ to one to facilitate the rest of our presentation. Now the H_∞ control problem can be stated as follows:

Find the necessary and sufficient conditions for the existence of an H_∞ controller. If one exists, obtain an explicit formula for all such controllers

This problem has been solved [1], [2]. Our objective is to solve this problem by a different approach based on the bounded real lemma given below, which turns out to give an explicit parameterization of all low-order H_∞ controllers in terms of a physically meaningful quantity. To this end, let us state the bounded real lemma modified for controller synthesis. We need the following definitions

$$Q := A_{cl} X_{cl} + X_{cl} A_{cl}' + B_{cl} B_{cl}' + (X_{cl} C_{cl}' + B_{cl} D_{cl}') R^{-1} (X_{cl} C_{cl}' + B_{cl} D_{cl}')' \quad (1)$$

$$R := I - D_{cl}' D_{cl} \quad (2)$$

$$\mathcal{C} := \{ (A_c, B_c, C_c, D_c) : R > 0 \text{ and } \exists X_{cl} \geq 0 \text{ s.t. } Q = 0 \} \quad (3)$$

Lemma 1: Let a plant (Σ_p) and a controller (Σ_c) be given. Suppose AP2) and AP3) hold. Then the following statements are equivalent.

i) The controller (Σ_c) is an H_∞ controller.

ii) The controller (Σ_c) is a minimal realization of some controller $(\tilde{\Sigma}_c) \in \mathcal{C}$.

Proof: Suppose i) holds. Then it is a standard fact (see, for example, [15] for a detailed proof and [16] for relaxation of the controllability assumption in [15]) that $(\Sigma_c) \in \mathcal{C}$. Hence ii) holds. Conversely, suppose ii) holds. Let matrices associated with $(\tilde{\Sigma}_c)$ be denoted with $\tilde{\cdot}$. If (A_c, B_c) is stabilizable, then $(\tilde{A}_c, \tilde{B}_c)$ is stabilizable and, by Lyapunov theory, \tilde{A}_c is asymptotically stable. In this case it is straightforward to show $\|\tilde{T}_{zw}\|_\infty \leq \gamma$. Hence, any minimal realization (Σ_c) of $(\tilde{\Sigma}_c)$ satisfies all four conditions C1)-C4), and we conclude i). Now, when (A_c, B_c) is not stabilizable, it can be shown by a coordinate change that a controllable realization of $(\tilde{\Sigma}_c)$, denoted by $(\tilde{\Sigma}_c)$, is such that $(\Sigma_c) \in \mathcal{C}$ and the corresponding pair $(\tilde{A}_c, \tilde{B}_c)$ is stabilizable. Then the same argument as above applies. This completes the proof.

If we view the (Riccati) equation $Q = 0$ as a Lyapunov equation with a forcing term quadratic in X_{cl} , a nonnegative definite solution X_{cl} can be used to prove closed-loop stability, while also defining the H_∞ performance. Hence we call X_{cl} an " H_∞ Lyapunov matrix." As mentioned in Section I, the "covariance control approach" to the H_∞ control problem taken in this note is the following. Consider the matrix equation $Q = 0$ as an algebraic problem to be solved for the controller parameters. In this case, solvability conditions characterize the set of all H_∞ Lyapunov matrices which can be "assigned" to the closed-loop system by some controller (Σ_c) . Then the general solution to $Q = 0$ provides an explicit formula for all controllers (Σ_c) which yields a specified X_{cl} as an H_∞ Lyapunov matrix. A controller constructed in this way may turn out to be such that (A_{cl}, B_{cl}) is not stabilizable, in which case, the usual bounded real lemma (e.g., [15]) cannot guarantee closed-loop stability. Lemma 1 stated above shows that, even for such a case, the closed-loop system is guaranteed to be internally stable.

Recall that the H_∞ Lyapunov matrix X_{cl} is related to the H_2 performance. As shown in [11], for the error signal $e := C_0 x$, the

H_2 norm of the transfer matrix $T_{e,w}$ from w to e is bounded above by a function of X_{cl} as follows

$$\|T_{e,w}\|_2^2 \leq J := \text{trace} C_0' X_p C_0'$$

where X_p is the 11-block ($n_p \times n_p$) of X_{cl} . The following lemma shows a necessary condition for an H_∞ controller to be "minimal order" in the H_2 related sense

Lemma 2 Given an n_c -th-order controller $(\Sigma_c) \in \mathcal{C}$. Let

$$X_{cl} := \begin{bmatrix} X_p & X_{pc} \\ X_{pc}' & X_c \end{bmatrix} \geq 0 \quad (4)$$

be any solution to $Q = 0$. Suppose $X_{pc}' X_{pc} \succ 0$. Then there exists a controller $(\hat{\Sigma}_c) \in \mathcal{C}$ of order $\hat{n}_c < n_c$ which yields the same H_2 performance bound J .

Proof Using the singular value decomposition of X_{pc} ($X_c^{1/2}$)⁺, define a matrix Ψ as

$$\Psi := \begin{bmatrix} I & 0 \\ 0 & V_1'(X_c^{1/2})^+ \end{bmatrix}, \quad X_{pc}(X_c^{1/2})^+ = U_1 \Sigma_1 V_1'$$

where $V_1 \in \mathbb{R}^{n_c \times n_c}$. Since $\text{rank}(X_{pc}) < n_c$, we have $n_c < n_c$. Noting that $X_{cl} \geq 0$ is equivalent to

$$X_c \geq 0, \quad X_p - X_{pc}' X_c^+ X_{pc} \geq 0, \quad (I - X_c X_c^+) X_c' = 0$$

computation of $Q := \Psi Q \Psi$ verifies that a controller (Σ_c) of order \hat{n}_c given by

$$\hat{A}_c = V_1'(X_c^{1/2})^+ A_c X_c^{1/2} V_1, \quad \hat{B}_c = V_1'(X_c^{1/2})^+ B_c,$$

$$C_c = C_c X_c^{1/2} V_1, \quad D_c = D_c$$

satisfies $Q = 0$ with

$$X_{cl} = \begin{bmatrix} X_p & U_1 \Sigma_1 \\ \Sigma_1 U_1' & I \end{bmatrix} \geq 0 \quad (5)$$

Since the 11-block of X_{cl} is identical to that of X_{cl} , the controller $(\hat{\Sigma}_c)$ yields the same H_2 performance bound.

By mathematical equivalency, if $X_{pc}' X_{pc} \succ 0$, then there exists a controller of lower order which has the same entropy at infinity when $C_1 = C_0$ [12]. By duality, the same statement can apply to the mixed H_2/H_∞ cost of Doyle *et al.* [13].

Lemmas 1 and 2 motivate us to define the following

Definition 2 A controller (Σ_c) is said to be an admissible H_∞ controller if $(\Sigma_c) \in \mathcal{C}$ and there exists a matrix $X_{cl} \geq 0$ which solves $Q = 0$ and satisfies $X_{pc}' X_{pc} \succ 0$.

Since we look for H_∞ controllers of low order, we will restrict our attention to admissible H_∞ controllers in the rest of this paper. It should be noted that the conditions $X_{pc}' X_{pc} \succ 0$ and $X_{cl} \geq 0$ imply that $X_c \succ 0$.

III ALL LOW ORDER H_∞ CONTROLLERS

This section provides a parameterization of all admissible H_∞ controllers in terms of the H_∞ Lyapunov matrix. Mathematical tools developed in covariance control theory are useful here [5]. To state the result, let \mathcal{L} be a subset of real symmetric matrices defined as follows. $X_{cl} \in \mathbb{R}^{(n_p+n_c) \times (n_p+n_c)}$ is an element of \mathcal{L} if

$$X_{cl} := \begin{bmatrix} X_p & X_{pc} \\ X_{pc}' & X_c \end{bmatrix} \geq 0, \quad X_{pc}' X_{pc} \succ 0$$

and there exist $L_{cl} \in \mathbb{R}^{n_p \times n_u}$, $L_1 \in \mathbb{R}^{n_p \times n_v}$ and $L_D \in \mathbb{R}^{n_u \times n_v}$ such that

$$A X_p + X_p A' + X_p C_1' C_1 X_p + B_1 B_1' - B_2 U^{-1} B_2' + L_{cl} L_{cl}' = 0 \quad (6)$$

$$A \bar{X}_p + \bar{X}_p A' + \bar{X}_p (C_1' C_1 - C_2' U^{-1} C_2) \bar{X}_p + B_1 B_1' + L_1 L_1' = 0 \quad (7)$$

$$(I - X_{pc}' X_{pc}^+) [L_{cl} L_{cl}' - \Gamma_{cl} R_{cl}^{-1} \Gamma_{cl}'] (I - X_{pc} X_{pc}^+) = 0, \quad (8)$$

$$\|D_{12} L_D D_{21}\| < 1 \quad (9)$$

where

$$\bar{X}_p = X_p - X_{pc}' X_c^{-1} X_{pc},$$

$$R_{cl} = U^{-1} - L_D V L_D', \quad \Gamma_{cl} = X_p C_2' L_D' + B_2 U^{-1}$$

The following theorem shows that \mathcal{L} is indeed the set of all H_∞ Lyapunov matrices X_{cl} which can be used to prove that a controller (Σ_c) is admissible. It also gives an explicit formula for all controllers which yield a given H_∞ Lyapunov matrix for the closed-loop system. A proof is given in the appendix. We shall need the following definitions

$$P = X_{cl} V^{-1}, \quad P_+ = X_{cl} V_1^{-1},$$

$$A = P_+ A P, \quad B_2 = P_+ B_2, \quad C_2 = C P \quad (10)$$

$$R_{cl} = V^{-1} - L_D' U L_D, \quad V_{cl} = B_2' L_D + X_{cl} C_2' V^{-1}$$

$$\Gamma_{cl} = \bar{X}_p C_2' L_D' + B_2 U^{-1}, \quad \Gamma_{cl} = B_2' L_D + X_{cl} C_2' V^{-1}$$

$$\Theta = I - X_{cl} X_{pc}^+, \quad \Gamma_{cl} = \Theta L_{cl}, \quad H_{cl} = \Theta L_{cl} R_{cl}^{-1/2}$$

$$F_{cl} = \Theta I_{cl}, \quad H_{cl} = \Theta I_{cl} R_{cl}^{-1/2}$$

Theorem 1 The following statements are equivalent

- i) There exists an admissible H_∞ controller
- ii) $\mathcal{L} \neq \emptyset$

In this case, let (Σ_c) be any controller. Then the following statements are equivalent

- iii) (Σ_c) is an admissible H_∞ controller
- iv) There exist $X_{cl} \in \mathcal{L}$ and orthogonal matrices Γ_{cl} and F_{cl} such that

$$A = \hat{A} + B_2 C_c - B_c C_2 + B_2 D C_2' + P_+ [\bar{X}_p C_1' C_1 - L_1 U_1 H_{cl}^{1/2} (B_c + N R_{cl}^{-1} D)' X_{pc}^+] P \quad (11)$$

$$B_c = P_+ (L_1 U_1 R_{cl}^{1/2} + \Gamma_{cl}),$$

$$C_c = (L_{cl} U_{cl} R_{cl}^{1/2} - \Gamma_{cl}')' (X_{cl}^+)', \quad D_c = L_D$$

where L_D satisfies (9) and

$$N := X_{cl} C_{cl}' + X_{pc}' C_2' D_{cl}' + B_{cl}' V D_{cl}'$$

$$U_{cl} = E_{cl}^+ H_{cl} + N_{FU} F_{cl} N_{HU}', \quad U_{cl} = -E_{cl}^+ H_{cl} + N_{LV} F_{cl} N_{HV}' \quad (12)$$

where columns of N_{FU} , N_{HU} , N_{LV} , and N_{HV} are orthonormal bases for the null spaces of E_{cl} , H_{cl} , E_{cl} , and H_{cl} , respectively.

Theorem 1 shows the observer-based structure of all (low order) H_∞ controllers. In particular, if we consider strictly proper controllers ($D_c = 0$), (11) yields

$$\dot{x}_c = \hat{A}x_c + \hat{B}_2 u + B_c(y - \hat{C}_2 x_c) + \hat{w},$$

$$u = C_c x_c, \quad \hat{w} := P_+(\bar{X}_p C_1' C_1 - L_1 U_V V^{-1/2} B_c' X_{pc}^+) P_{X_c}$$

where $(\hat{A}, \hat{B}_2, \hat{C}_2)$ are defined in (10). Thus any low order strictly proper admissible H_∞ controller has the structure of state-estimator plus estimated-state feedback for the "reduced-order model" $(\hat{A}, \hat{B}_2, \hat{C}_2)$ of the original plant, obtained by keeping only the states belonging to the range space of X_{pc} (or $X_p - \bar{X}_p$). As in the full-order case (see [1]), a low-order H_∞ estimator also has an extra term \hat{w} , which can be thought of loosely as an estimate for the worst case L_2 disturbance.

Note that the two Riccati equations (6) and (7) are the familiar ones which describe the existence conditions for an H_∞ controller [2] if i) X_p is invertible so that (6) can be multiplied by X_p^{-1} from both sides, and ii) the free matrices L_c and L_1 are set to zero. Since the controller order is fixed (n_c is specified by the dimension of $X_{c,t}$), we have another condition (8). The familiar spectral radius condition in [1] is implicitly embedded in the set \mathcal{L} : by the definition of X_p and the positiveness of $X_{c,t}$, we require $X_p \geq \bar{X}_p \geq 0$, which is equivalent to $\rho(X_p^{-1} \bar{X}_p) \leq 1$ when $X_p > 0$. Notice that the third condition (8) disappears if we consider the full-order ($n_c = n_p$) controller case since X_{pc} becomes square and invertible. Thus a full-order H_∞ controller can be designed by solving two Riccati equations. The more detailed exploration of the full-order case, including explicit upper bounds on the plant state covariance and the control input covariance in terms of the H_∞ Lyapunov matrix, can be found in the conference version of this note [17]. For the general case ($n_c < n_p$), however, finding $X_{c,t} \geq 0$ that satisfies all three conditions (6)–(8) does not seem to be tractable. Homotopic continuation methods [18] may be useful to solve the coupled Riccati equations.

IV. CONCLUSION

The set of all H_∞ controllers of order equal to or less than the plant is parameterized in terms of the H_∞ Lyapunov matrix, which carries many system properties such as H_2 -related performances and the system entropy. The H_∞ Lyapunov matrices, which can be "assigned" to the closed-loop system by some controller, are characterized by the two Riccati equations and the coupling condition in Theorem 1, which are numerically nontrivial to solve for the general low-order controller case. Hence, the computational issue remains open for further research. The coupling condition disappears for the full-order case, and thus the computational problem becomes much easier as in the literature [1]. Our novel results show that any low-order H_∞ controller has observer-based structure. This has been shown [1] only for the full-order (central) H_∞ controller. The observer part of the low-order H_∞ controller estimates the states of the "reduced-order model" of the original plant obtained by some model reduction methods based on a projection formulation.

APPENDIX

Proof of Theorem 1: Recall that all H_∞ controllers are characterized by $Q = 0$ and $R > 0$ in (1) and (2). Note that R is positive definite if and only if $\|D_{12} D_{21}\| < 1$. So we consider the following mathematical problem: Given $D_c (= L_{11})$ such that the above norm bound constraint is satisfied, find necessary and sufficient conditions for the solvability of $Q = 0$ for (A, B, C) . The proof begins with necessity.

Necessity: Suppose (Σ_c) is an admissible H_∞ controller. Then $Q = 0$ holds for some $X_{c,t} \geq 0$, or equivalently

$$\begin{bmatrix} I & -X_{pc} X_{c,t}^{-1} \\ 0 & I \end{bmatrix} Q \begin{bmatrix} I & 0 \\ -X_{c,t}^{-1} X_{pc}' & I \end{bmatrix} = 0$$

whose partitioned blocks are given by (c.f. [7])

$$A \bar{X}_p + \bar{X}_p A' + \bar{X}_p (C_1' C_1 - C_2' V^{-1} C_2) \bar{X}_p + B_1 B_1' + (X_{pc}' X_{c,t}^{-1} B_c - \bar{\Gamma}_1) R_V^{-1} (X_{pc}' X_{c,t}^{-1} B_c - \bar{\Gamma}_1)' = 0, \quad (13)$$

$$(A + \bar{X}_p C_1' C_1) X_{pc} + \bar{X}_p C_2' B_c' - X_{pc}' X_{c,t}^{-1} A_c X_c + \bar{\Gamma}_1' R_V^{-1} N' - X_{pc}' X_{c,t}^{-1} B_c (C_2 X_{pc} + V B_c' + V' L_D R_V^{-1} N') = 0, \quad (14)$$

$$A_c X_c + X_c A_c' + X_{pc}' (C_1' C_1 - C_2' V^{-1} C_2) X_{pc} + N R_V^{-1} N' + (B_c + X_{pc}' C_2' V^{-1}) V' (B_c + X_{pc}' C_2' V^{-1})' = 0. \quad (15)$$

Also, it is necessary that 11-block of Q is equal to zero, i.e.,

$$A X_p + X_p A' + X_p C_1' C_1 X_p + B_1 B_1' - B_2 U^{-1} B_2' + (X_{pc}' C_c' + \Gamma_{11}) R_{11}^{-1} (X_{pc}' C_c' + \Gamma_{11})' = 0. \quad (16)$$

Since the last term in (16) is nonnegative definite with rank $\leq n_u$, it is necessary that (6) holds for some $L_{11} \in \mathbb{R}^{n_p \times n_u}$ satisfying

$$L_{11} L_{11}' = (X_{pc}' C_c' + \Gamma_{11}) R_{11}^{-1} (X_{pc}' C_c' + \Gamma_{11})'. \quad (17)$$

Similarly for (13), it is necessary that (7) holds for some $L_1 \in \mathbb{R}^{n_p \times n_v}$ satisfying

$$L_1 L_1' = (X_{pc}' X_{c,t}^{-1} B_c - \bar{\Gamma}_1) R_V^{-1} (X_{pc}' X_{c,t}^{-1} B_c - \bar{\Gamma}_1)'. \quad (18)$$

Now, applying the result (Appendix C) of [5], there exists a C_c satisfying (17) if and only if the equality in (8) holds, and all solutions C_c are given by

$$C_c = (L_{11}' U_{11}' R_{11}^{1/2} - \Gamma_{11}')' (X_{pc}^+)' \quad (19)$$

where U_{11} is defined by (12) with F_{11} being an arbitrary orthogonal matrix. Thus we have established the necessity of (6)–(9).

Sufficiency: Now, suppose conditions (6)–(9) hold. We need to show the existence of a controller which satisfies (13)–(15). In the necessity part of the proof, we have shown that C_c given by (19) satisfies (16) under supposed conditions. To prove sufficiency, we will construct B_c and A_c satisfying (13)–(15) together with C_c given by (19).

Using the result of [5] again, there exists a B_c satisfying (18) if and only if

$$(I - X_{pc}' X_{pc}^+) [L_1 L_1' - \bar{\Gamma}_1 R_V^{-1} \bar{\Gamma}_1'] (I - X_{pc}' X_{pc}^+) = 0 \quad (20)$$

holds and all solutions B_c are given by

$$B_c = X_c X_{pc}^+ (L_1 U_{11}' R_V^{1/2} + \bar{\Gamma}_1) \quad (21)$$

where U_{11} is defined by (12) with F_{11} being an arbitrary orthogonal matrix. It can be verified that the conditions (6)–(9) imply that (20) always holds. Thus, B_c given by (21) solves (13).

Next, note that (14) is solvable for A_c if and only if $(I - X_{pc}' X_{pc}^+) \hat{Q}_{12} = 0$ holds where \hat{Q}_{12} is the left hand side of (14). In this case, the unique solution A_c is given by

$$A_c = X_c X_{pc}^+ \hat{Q}_{12} X_c^{-1}. \quad (22)$$

Now we claim that the above solvability condition is always satisfied by the choice of B_c and C_c given by (21) and (19), respectively. To

prove this, it is sufficient to show that

$$\bar{Q}_{12} := (I - X_{pc} X_{pc}^+) \bar{Q}_{12} X_c^{-1} X_{pc}' = 0 \quad (23)$$

since $X_c > 0$ and $X_{pc}' X_{pc} > 0$. Substituting (19) and (21) into (23), and using (12), after some manipulations, we obtain $\bar{Q}_{12} = 0$. Thus A_c , B_c , and C_c given by (22), (21), and (19) solve (13) and (14) provided (6)–(9) holds.

Finally, straightforward algebraic manipulations show that (15) also holds. The basic steps are the following. First solve (14) for $X_{pc} X_c^{-1} A_c X_c$ and then substitute into

$$\bar{Q}_{22} = X_{pc} X_c^{-1} Q_{22} X_c^{-1} X_{pc}'$$

where Q_{22} denotes the left hand side of (15). Using (7)–(9) and (17)–(18), after some manipulations, it can be verified that $\bar{Q}_{22} = 0$, which implies (15). This completes the proof. Q E D

REFERENCES

- [1] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, "State space solutions to standard H_2 and H_∞ control problems," *IEEE Trans Automat Contr*, vol. 34, pp. 831–847, August 1989.
- [2] K. Glover and J. Doyle, "State-space formulae for all stabilizing controllers that satisfy an H_∞ norm bound and relations to risk sensitivity," *Sys Contr Lett*, vol. 11, pp. 167–172, 1988.
- [3] A. Holz and R. E. Skelton, "Covariance control theory," *Int J Contr*, vol. 46, pp. 13–32, 1987.
- [4] R. E. Skelton and T. Iwasaki, "Lyapunov and covariance controllers," in *Proc Amer Contr Conf*, 1992, pp. 2861–2865.
- [5] —, "Lyapunov and covariance controllers," *Int J Contr*, vol. 57, pp. 519–536, 1993.
- [6] K. Yasuda, R. F. Skelton, and K. M. Grigoriadis, "Covariance controllers: A new parametrization of the class of all stabilizing controllers," *Automatica*, vol. 29, pp. 785–788, 1993.
- [7] P. M. Gahinet, "A new representation of H_∞ suboptimal controllers," in *Proc Amer Contr Conf*, 1992, pp. 2240–2244.
- [8] I. R. Petersen, "Disturbance attenuation and H_∞ optimization: A design method based on the algebraic Riccati equation," *IEEE Trans Automat Contr*, vol. AC-32, pp. 427–429, May 1987.
- [9] M. Sampaio, T. Mita, and M. Nakamichi, "An algebraic approach to H_∞ output feedback control problems," *Sys Contr Lett*, vol. 14, pp. 13–24, 1990.
- [10] K. Zhou and P. P. Khargonekar, "An algebraic Riccati equation approach to H_∞ optimization," *Sys Contr Lett*, vol. 11, pp. 85–92, 1988.
- [11] D. S. Bernstein and W. M. Haddad, "LQG control with an H_∞ performance bound: A Riccati equation approach," *IEEE Trans Automat Contr*, vol. 34, pp. 293–305, Mar 1989.
- [12] D. Mustafa, "Relations between maximum-entropy/ H_∞ control and combined H_∞ /LQG control," *Sys Contr Lett*, vol. 12, pp. 193–203, 1989.
- [13] J. C. Doyle, K. Zhou, and B. Bodenheimer, "Optimal control with mixed H_2 and H_∞ performance objectives," in *Proc Amer Contr Conf*, 1989, pp. 2065–2070.
- [14] J. Stoustrup, R. E. Skelton, and T. Iwasaki, "Mixed H_2/H_∞ state feedback control with improved covariance bounds," in *Proc IFAC World Congress*, Sydney, Australia, July, 1993.
- [15] J. C. Willems, "Least squares stationary optimal control and the algebraic Riccati equation," *IEEE Trans Automat Contr*, vol. AC-16, pp. 621–634, 1971.
- [16] C. Scherer, " H_∞ -control by state feedback: An iterative algorithm and characterization of high-gain occurrence," *Sys Contr Lett*, vol. 12, pp. 383–391, 1989.
- [17] T. Iwasaki and R. E. Skelton, "All low order H_∞ controllers with covariance upper bound," in *Proc Amer Contr Conf*, 1993, pp. 2180–2184.
- [18] S. Richter, "A homotopy algorithm for solving the optimal projection equations for fixed-order dynamic compensation: Existence, convergence and global optimality," in *Proc Amer Contr Conf*, vol. 3, 1987, pp. 1527–1531.

Intrinsic Difficulties in Using the Doubly-Infinite Time Axis for Input-Output Control Theory

Tryphon T. Georgiou and Malcolm C. Smith

Abstract—We point out that the natural definitions of stability and causality in input-output control theory lead to certain inconsistencies when inputs and outputs are allowed to have support on the doubly-infinite time-axis. In particular, linear time-invariant systems with right-half plane poles cannot be considered to be both causal and stabilizable. In contrast, there is no such conflict when the semi-infinite time axis is used.

I. DISCUSSION

Consider two systems P_i ($i = 1, 2$) defined in terms of convolution representations

$$y(t) = \int_{-\infty}^{\infty} h_i(t - \tau) u(\tau) d\tau = h_i * u$$

where $h_1(t) = e^t$ for $t \geq 0$ and zero otherwise, and $h_2(t) = -e^t$ for $t \leq 0$ and zero otherwise. Both systems have Laplace transfer functions $1/(s-1)$, but with differing regions of convergence. The first system is unstable and causal, and the second is stable and noncausal (in fact anticausal) according to the usual definitions. To view these systems abstractly in the input-output setting (or in the behavioral framework of [3]) we need to work out the system trajectories for signals in some function space, say \mathcal{L}_2 . This amounts to finding the graph, i.e., the set of input-output pairs in \mathcal{L}_2 . We mention that an explicit working out of such an approach for the semi-infinite time-axis is given in [2]. The remarks below are concerned with the case of the doubly-infinite time-axis.

We consider first P_2 . From [1 p. 158] $f \in \mathcal{L}_1$ and $g \in \mathcal{L}_1$ implies $f * g \in \mathcal{L}_1$ and

$$\|f * g\|_1 \leq \|f\|_1 \|g\|_1$$

Thus, P_2 can be alternatively represented by the following graph on $\mathcal{L}_2(-\infty, \infty)$

$$\mathcal{G}_{P_2}(-\infty, \infty) = \left(\begin{array}{c} 1 \\ \frac{1}{s-1} \end{array} \right) \mathcal{L}_2(-j\infty, j\infty) \quad (1)$$

after transforming to the frequency domain. (As usual, \mathcal{L}_2 denotes the Fourier transform which maps $\mathcal{L}_2(-\infty, \infty)$ isometrically and isomorphically onto $\mathcal{L}_2(-j\infty, j\infty)$ and $\mathcal{L}_2[0, \infty)$ onto the Hardy space \mathcal{H}_2 of the right-half plane [1, Chapter 19]. Functions in \mathcal{H}_2 can be extended analytically into the right-half plane by replacing $j\omega$ by s .)

Now consider P_1 . We will first restrict the inputs to the space $\mathcal{L}_2[0, \infty)$. Since

$$y(t) = \int_0^t e^{t-\tau} u(\tau) d\tau = e^t \int_0^t e^{-\tau} u(\tau) d\tau$$

a necessary condition for $y(t) \in \mathcal{L}_2[0, \infty)$ is that

$$\int_0^t e^{-\tau} u(\tau) d\tau \rightarrow 0$$

Manuscript received March 29, 1994; revised June 14, 1994. This work was supported in part by the NSF and the AFOSR.

T. T. Georgiou is with the Department of Electrical Engineering, University of Minnesota, Minneapolis, MN 55455 USA.

M. C. Smith is with the University of Cambridge, Department of Engineering, Cambridge CB2 1PZ, U.K.
IEEE Log Number 9407262

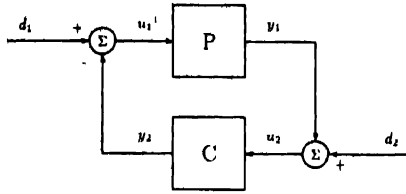


Fig. 1. Standard feedback configuration.

as $t \rightarrow \infty$, which is equivalent to $\langle e^{-t}, u \rangle = 0$. In fact, this is also sufficient since then

$$y(t) = -e^t \int_t^\infty e^{-\tau} u(\tau) d\tau = h_2 * u.$$

Thus the domain of P_1 in $\mathcal{L}_2[0, \infty)$ is equal to the orthogonal complement of $e^{-t}\mathbb{C}$ and, moreover, P_1 coincides with P_2 on this domain. After taking transforms, the orthogonal complement of $1/(s+1)\mathbb{C}$ is $(s-1)/(s+1)\mathcal{H}_2$. Therefore the graph of P_1 restricted to signals with support on $[0, \infty)$ becomes (in the frequency domain)

$$\hat{G}P_1|_{[0, \infty)} = \begin{pmatrix} \frac{s-1}{s+1} \\ 1 \end{pmatrix} \mathcal{H}_2 =: GH_2.$$

Since P_1 is shift invariant, the graph of P_1 on $\mathcal{L}_2(-\infty, \infty)$ contains (again, expressed in the frequency domain) the infinite union

$$\bigcup_{t \geq 0} e^{st} GH_2. \quad (2)$$

The graph of P_1 on $\mathcal{L}_2(-\infty, \infty)$ is actually slightly bigger than (2) (see below) but we will not need it for the next part of the discussion.

We now turn our attention to the requirement of stabilizability in the feedback configuration of Fig. 1. Our definition of stability is the usual one: for all external \mathcal{L}_2 disturbance inputs, there must be solutions of the feedback equations in \mathcal{L}_2 so that the closed-loop operators are norm bounded. If P_1 is stabilized by some compensator C on $\mathcal{L}_2(-\infty, \infty)$, then it turns out that the graph of P_1 must be closed. (This is a standard argument which proceeds as follows. Take a Cauchy sequence in the graph of P_1 and set this equal to $\begin{pmatrix} d_1 \\ d_2 \end{pmatrix}$. Since by assumption the feedback equations have unique solutions they must be given by: $u_1 = d_1$, $y_1 = d_2$, $u_2 = 0$, $y_2 = 0$. Now take the limit of the Cauchy sequence at the external inputs. Since the closed-loop operators are bounded, this pair of signals must also appear at u_1, y_1 .) Thus, the graph of P_1 must contain the closure of (2). But the closure of the infinite union (2) equals the graph of P_2 ! Moreover, since P_2 is an anticausal operator, then there are $\mathcal{L}_2(-\infty, \infty)$ input-output pairs which satisfy the convolution representation for P_2 but not for P_1 . Such a pair is: $u(t) = e^{-t}(t \geq 0)$, 0 ($t < 0$) and $y(t) = -e^{-|t|}/2$, which equals

$$\begin{pmatrix} \hat{u} \\ \hat{y} \end{pmatrix} = \begin{pmatrix} \frac{1}{s+1} \\ \frac{1}{(s-1)(s+1)} \end{pmatrix} \quad (3)$$

in the transform domain. Thus, it seems incorrect to close the graph of P_1 and use the ordinary definition of input-output stability. On the other hand, stabilizability of P_1 would require that any pair of possible disturbance signals d_1, d_2 could act on the feedback system in Fig. 1 and produce a bounded response. If, however, a disturbance

$$\begin{pmatrix} \hat{d}_1 \\ \hat{d}_2 \end{pmatrix} = \begin{pmatrix} \hat{u} \\ \hat{y} \end{pmatrix}$$

with \hat{u}, \hat{y} as in (3), is allowed to act on the input ports in Fig. 1, then there exists no solution which is consistent with the feedback equations and the integral representation of P_1 .

Of the following possible remedies, none seem to be satisfactory:

- 1) To consider P_1 to be nonstabilizable on $\mathcal{L}_2(-\infty, \infty)$, and to restrict attention to open-loop stable systems only.
- 2) To seek an alternative definition of closed-loop stability for the $\mathcal{L}_2(-\infty, \infty)$ case which would agree with the common belief that P_1 is stabilized by proportional negative feedback of gain greater than one.
- 3) To identify the systems P_1, P_2 .

It should be noted that the use of extended spaces does not improve the situation, since the difficulty is to determine the correct behavior for signals in $\mathcal{L}_2(-\infty, \infty)$.

Option (1) is a correct conclusion based on existing definitions. This would mean, however, that the doubly-infinite set-up is of limited interest for control purposes, since systems which in the usual sense are "open-loop unstable" would have to be excluded. Option (2) does not seem to provide any satisfactory alternatives. It is, of course, possible to restrict attention to the subspace of \mathcal{L}_2 consisting of functions which have support on some interval $[T, \infty)$ for some arbitrary finite T . This would treat the graph of P_1 precisely as (2) and would work with driving signals in $\bigcup_{T \geq 0} e^{sT} \mathcal{H}_2$, which is not a closed subspace of $\mathcal{L}_2^2(-j\infty, j\infty)$. This appears to be a rather cosmetic solution, however, which more or less amounts to working on the half line. A more natural avenue would be to consider the actual $\mathcal{L}_2(-\infty, \infty)$ graph of the convolution operator P_1 . Similar reasoning to the above shows that this is the same as the graph of P_2 , but with the restriction that the inputs satisfy

$$\int_{-\infty}^{\infty} e^{-\tau} u(\tau) d\tau = 0.$$

Again, this means that the graph P_1 is not a closed subspace of $\mathcal{L}_2^2(-\infty, \infty)$ and the problem is then to find a suitable subspace for the external driving signals. The option of trying to work on some subspace of $\mathcal{L}_2(-\infty, \infty)$ with signals which "decay sufficiently fast" towards minus infinity again does not seem to offer a satisfying resolution. Option (3), although unsatisfactory, is perhaps not as outrageous as would first appear. If we consider the system represented by the differential equation

$$\dot{y} = y + u$$

we can reproduce the trajectories of P_1 by solving this equation forwards in time and those of P_2 by solving it backwards. This more or less amounts to abandoning any notion of causality (compare with the need to consider the anticausal trajectory (3) as a valid input-output pair for P_1). This is not a natural option, however, if the direction of time is well defined (which is a basic assumption if we consider questions of control.)

II. CONCLUSION

The purpose of this note has been to point out certain features of input-output control theory on the doubly infinite time axis which appear intrinsically unsatisfactory. The difficulties are not of a mathematical nature—a consistent picture is obtained with a variety of definitions. The problem lies in trying to escape from conclusions which limit the engineering relevance of the theory, e.g., among the causal systems, only the stable ones are stabilizable. If the definitions lead to such a conclusion, then it would not seem worthwhile to develop an elaborate theory of stabilization in that context. In contrast we remark that input-output systems theory on the doubly infinite time axis does have many important uses, e.g., in discussions of fundamental limitations in filtering imposed by causality conditions or in system approximation using Hankel operators.

REFERENCES

- [1] W. Rudin, *Real and Complex Analysis*, 2nd ed. New York: McGraw-Hill, 1982.
- [2] T. T. Georgiou and M. C. Smith, "Graphs, causality and stabilizability: Linear, shift-invariant systems on $L_2[0, \infty)$," *Math. Contr., Signals Syst.*, vol. 6, no. 3, pp. 195–223, 1993.
- [3] J. C. Willems, "Paradigms and puzzles in the theory of dynamical systems," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 259–294, 1991.

Regional Observability of a Thermal Process

A. El Jai, E. Zerrik, M. C. Simon, and M. Amouroux

Abstract—The realization of a ceramic protector covering a subregion of a substrate and using chemical vapor deposition (cvd) techniques needs the temperature field in the deposition zone to be controlled. A feedback control of this temperature is based on the knowledge of the temperature in the considered subregion. The difficulty occurs because measurements can only be obtained out of the deposition zone. The purpose of this paper is to give an original approach for the reconstruction of the state in the deposition subregion.

I. MOTIVATION

Thermal treatment is a usual technique to elaborate materials. The quality of the produced material depends on the control of the temperature field. The purpose of the treatment may be either the improvement of mechanical properties of surfaces (to be reinforced by metallic or by ceramic substrate) or the prevention against corrosion or oxidation.

In the considered problem we try to realize a ceramic protector using chemical vapor deposition (cvd) techniques [7]. Only a part of the substrate is concerned with the deposition. Thermal and mechanical qualities of the obtained material depend on the homogeneity of the temperature field of the material. The control of the temperature field in the concerned zone is determined by observations outside the concerned zone. Such a problem can be solved with using a regional observability concept [4].

The aim of this paper is to give an extension of parabolic systems observability connected to the above problem. If the system is defined in a space domain Ω , we are concerned with the state observation not in the whole domain Ω but only in a given subregion $\omega \subset \Omega$. This situation occurs in many practical applications where the knowledge of the state is necessary only in a critical subregion. The measurements are given by means of a finite number of sensors which may be pointwise or zone and located inside Ω or in its boundary $\partial\Omega = \Gamma$. Let y be the state of a system (S) with a state space $X = L^2(\Omega)$, and assume that a state y_0 at a given time t_0 is unknown. Suppose now that measurements are given by means of an output $z \in Z$ (depending on the number and the structure of the sensors, the measurement interval, ...). The studied problem concerns the reconstruction of the state y_0 on the subregion ω .

The next section is devoted to the regional reconstruction of the state, then we develop a numerical approach, and in the last section the results are applied to the considered thermal process.

Manuscript received February 2, 1993; revised December 8, 1993 and June 17, 1994.

The authors are with IMP/CNRS, University of Perpignan, 52, Avenue de Villeneuve, F-66860 Perpignan Cedex, France.

IEEE Log Number 9407231.

II. REGIONAL OBSERVATION PROBLEM

A. Problem Statement

Let

- Ω be a regular bounded open set of \mathbb{R}^n , ($n = 1, 2$ or 3) with boundary $\Gamma = \partial\Omega$.
- ω be a nonempty given subregion of Ω .
- $[0, t_f]$ with $t_f > 0$ a time measurement interval, we denote $Q = \Omega \times]0, t_f[$ and $\Sigma = \Gamma \times]0, t_f[$.
- A be a linear differential operator defined by

$$A = a_0 - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial}{\partial x_j} \right) \quad (2.1)$$

where $a_0, a_{ij} \in \mathcal{D}'(\bar{\Omega} \times [0, t_f])$, the space of C^∞ functions with compact support in $\mathcal{D}'(\bar{\Omega} \times [0, t_f])$. We assume that A is elliptic; A^* is the adjoint operator of A .

We consider the system described by the state-space equation

$$\begin{cases} y' + Ay = 0 & \text{in } Q \\ y = 0 & \text{in } \Sigma \\ y(0) = y_0 \text{ supposed unknown} & \text{in } \Omega \end{cases} \quad (2.2)$$

and the output function

$$z(t) = C'y(t) \quad (2.3)$$

where

$$C': y \in L^2(0, t_f; L^2(\Omega)) \longrightarrow z \in L^2(0, t_f; Z). \quad (2.4)$$

Equation (2.3) gives measurements of the state of the system. These measurements can be obtained by pointwise or zone sensors which may be located in Ω or in Γ .

Let us recall that a sensor is defined by a couple (D, f) (see [1]) where:

- i) $D \subset \Omega$ is the support of the sensor in the zone case, while $D = \{b\}$ in the pointwise case; and
- ii) $f \in L^2(D)$ is the spatial distribution of the sensor in the zone case, while $f = \delta_b$ in the pointwise case (δ_b is the Dirac mass in b).

Let

$$y_0 = \begin{cases} y_0^1 & \omega \text{ to be estimated} \\ y_0^2 & \Omega \setminus \omega \text{ undesired} \end{cases} \quad (2.5)$$

The problem consists of reconstructing y_0^1 with the knowledge of (2.2) and (2.3). From (2.2)–(2.3) we can write $z = Ky_0$ where K is as an operator: $L^2(\Omega) \rightarrow L^2(0, t_f; Z)$. We recall the following definitions.

Definition 2.1:

- 1) System (2.2) with output (2.3) is weakly observable if $\ker K = \{0\}$. The sensor is then said to be strategic [2].
- 2) System (2.2) with output (2.3) is weakly regionally observable in $\omega \subset \Omega$ if $\ker K|_{\omega} = \{0\}$. In this case the sensor is said to be ω -strategic with

$$\begin{aligned} \chi_\omega: L^2(\Omega) &\longrightarrow L^2(\omega) \\ \chi_\omega z &= z|_\omega \end{aligned} \quad (2.6)$$

($z|_\omega$ is the restriction of z to ω .) the adjoint χ_ω^* of χ_ω

$$\chi_\omega^*: L^2(\omega) \longrightarrow L^2(\Omega) \quad (2.7)$$

is defined by

$$\omega f = \begin{cases} f & \omega \\ 0 & \Omega \setminus \omega \end{cases} \quad (2.8)$$

The concept of regional analysis has been recently developed by Jai and Zerrik, and various papers have been devoted to this new approach of distributed parameter systems [3], [4].

B. Reconstruction Method: Case of a Pointwise Sensor

In this case, system (2.2) is observed via the output

$$z(t) = y(b, t) \quad (2.9)$$

where $b \in \Omega$ is the sensor location. Consider the set

$$G = \{g \in L^2(\Omega) \mid g = 0 \text{ in } \Omega \setminus \omega\} \quad (2.10)$$

and

$$\tilde{G} = \{\tilde{g} \in L^2(\Omega) \mid \tilde{g} = 0 \text{ in } \omega\} \quad (2.11)$$

and denote $\langle \cdot, \cdot \rangle$ the $L^2(\Omega)$ scalar product. It is then easy to see that for $g \in G$ and $\tilde{g} \in \tilde{G}$

$$\langle g, \tilde{g} \rangle = \int_{\Omega} g \tilde{g} dx = \int_{\omega} g \tilde{g} dx + \int_{\Omega \setminus \omega} g \tilde{g} dx = 0. \quad (2.12)$$

For $\varphi^0 \in G$, consider the system

$$\begin{cases} \varphi' + A\varphi = 0 & Q \\ \varphi = 0 & \Sigma \\ \varphi(0) = \varphi^0 & \Omega \end{cases} \quad (2.13)$$

which has a unique solution $\varphi \in L^2(Q)$. Moreover φ is continuous on Q . Then we can consider the operator $\|\cdot\|_G$ defined by

$$\varphi^0 \in G \longrightarrow \|\varphi^0\|_G^2 = \int_0^t \varphi^2(b, t) dt \quad (2.14)$$

which is a semi-norm on G . Consider now the system

$$\begin{cases} -\psi' + A^* \psi = -\varphi(b, t) \delta(x - b) & Q \\ \psi = 0 & \Sigma \\ \psi(t_f) = 0 & \Omega \end{cases} \quad (2.15)$$

For $\varphi^0 \in G$, (2.13) gives $\varphi(b, t)$ and with (2.15) we can obtain a corresponding $\psi(0)$.

Let Λ be the operator defined by

$$\Lambda \varphi^0 = P_{G^\perp}(\psi(0)) \quad (2.16)$$

where P_{G^\perp} is the projection on \tilde{G}^\perp (\tilde{G}^\perp may be identified to the dual G' of G). Let \bar{Z} be defined by the solution of

$$\begin{cases} -\bar{Z}' + A^* \bar{Z} = -z(t) \delta(x - b) & Q \\ \bar{Z} = 0 & \Sigma \\ \bar{Z}(t_f) = 0 & \Omega \end{cases} \quad (2.17)$$

Finally for a convenient choice of φ^0 in G , (i.e., such that $\varphi(b, t) = z(t)$), system (2.15) is the adjoint of the observed system (2.2)–(2.9) and hence, the observation problem in the subregion ω will be naturally solved by the equation

$$\Lambda \varphi^0 = P_{\tilde{G}^\perp}(\bar{Z}(0)). \quad (2.18)$$

The considered approach is derived from Lions Hilbert uniqueness method [5] which has been developed for the usual controllability concept but here the context is somewhat different. We have the following result.

Proposition 2.2: If the sensor (b, h_b) is ω -strategic then (2.18) has a unique solution $\varphi^0 \in G$ which corresponds to the regional state y_0^1 to be observed in ω .

Proof:

- 1) Firstly we show that if the sensor (b, h_b) is ω -strategic then (2.14) defines a norm on G . For $\varphi^0 \in G$, $\|\varphi^0\|_G = 0 \iff \varphi(b, t) = 0$ for all $t \in [0, t_f]$. As the system is autonomous we have $\varphi^0 \in \ker K_{\omega}^*$ and as the sensor (b, h_b) is ω -strategic then $\varphi^0 = 0$ in Ω .
- 2) The uniqueness of (2.18) is immediate if Λ is an isomorphism from $G \longrightarrow \tilde{G}^\perp$.

We have

$$\begin{aligned} \langle \Lambda \varphi^0, \varphi^0 \rangle &= \langle P_{G^\perp}(\psi(0)), \varphi^0 \rangle \\ &= \langle \psi(0), \varphi^0 \rangle \end{aligned}$$

and a formal use of Green's formula gives, by multiplying (2.15) by φ and integrating by parts

$$\begin{aligned} \langle \psi(t_f), \varphi(t_f) \rangle &= \langle \psi(0), \varphi^0 \rangle \\ &\quad - \int_Q \psi(-\varphi' + A^* \varphi) dQ \\ &\quad + \int_{\Sigma} \varphi \frac{\partial \psi}{\partial \nu} d\Sigma - \int_{\Sigma} \psi \frac{\partial \varphi}{\partial \nu} d\Sigma \\ &= - \int_0^{t_f} \varphi(b, t)^2 dt \end{aligned}$$

where ν is the outward normal derivative to $\partial\Omega$ and, for the operator A defined in (2.1)

$$\frac{\partial f}{\partial \nu} = \sum_{i=1}^n a_{ij}(x, t) \frac{\partial f}{\partial x_j} \cos(\nu, x_i).$$

From boundary and initial conditions we have

$$\begin{aligned} \langle \Lambda \varphi^0, \varphi^0 \rangle &= \langle P_{G^\perp}(\psi(0)), \varphi^0 \rangle \\ &= \|\varphi^0\|_G^2. \end{aligned}$$

One deduces that Λ is an isomorphism from G to \tilde{G}^\perp (all this can be made completely rigorous without too much difficulty). Finally (2.18) has a unique solution $\varphi^0 = y_0^1$, which ensures the observation of the initial state in the subregion ω .

Remarks:

- 1) In [4] we show that (b, h_b) is ω -strategic $\iff \varphi_i(b) \neq 0$ for all i , where (φ_i) is the set of eigenfunctions of A , with Dirichlet boundary conditions, in $L^2(\omega)$ associated to the eigenvalues λ_i .
- 2) These results may be easily extended to other kinds of measurements [4].

III. NUMERICAL APPROACH

The numerical approach may be achieved very easily when one can calculate the eigenfunctions of the system: this is the case of invariant systems and will be developed in the next subsection. In the general case an adapted technique is given and detailed later.

We have seen that the solution of the regional observation problem is obtained by the solution of the equation

$$\Lambda \varphi^0 = P_{\tilde{G}^\perp}(\bar{Z}(0)) \varphi^0 \in G. \quad (3.1)$$

In the next sections we give an implementable approach for solving the above equation.

A. Invariant Systems Case

In this section we consider the case where the system is described by an equation with constant coefficients. The idea, in this case, is to calculate as shown below, in a suitable basis (φ_j) of $L^2(\Omega)$, the components $\Lambda_{i,j}$ of Λ .

We assume that A has in $L^2(\Omega)$ a complete set of eigenfunctions (φ_j) such that

$$\begin{cases} A\varphi_j = \lambda_j \varphi_j & \Omega \\ \varphi_j = 0 & \Gamma \\ \|\varphi_j\|^2 = 1 & \forall j \end{cases} \quad (3.2)$$

Without loss of generality we suppose that the eigenvalues λ_i of A are of multiplicity one. As the sensor is pointwise, we have

$$\langle \Lambda \varphi^0, \varphi^0 \rangle = \|\varphi^0\|_C^2 = \int_0^t \varphi(b, t)^2 dt \quad (3.3)$$

and with

$$\varphi(b, t) = \sum_{i=1}^{\infty} \langle \varphi^0, \varphi_i \rangle_{L^2(\omega)} \varphi_i(b) e^{\lambda_i t} \quad (3.4)$$

we have

$$\begin{aligned} \langle \Lambda \varphi^0, \varphi^0 \rangle &= \sum_{i,j} \langle \varphi^0, \varphi_i \rangle_{L^2(\omega)} \langle \varphi^0, \varphi_j \rangle_{L^2(\omega)} \\ &\quad \times \varphi_i(b) \varphi_j(b) \frac{e^{(\lambda_i + \lambda_j)t} - 1}{\lambda_i + \lambda_j}. \end{aligned} \quad (3.5)$$

One deduces

$$\Lambda = (\Lambda_{i,j}), \quad i, j = 1, \infty$$

with

$$\Lambda_{i,j} = \varphi_i(b) \varphi_j(b) \frac{e^{(\lambda_i + \lambda_j)t} - 1}{\lambda_i + \lambda_j} \quad (3.6)$$

hence, the regional observation problem is transformed in the linear system

$$\sum_{j=1}^M \Lambda_{i,j} \varphi_j^0 = \bar{Z}'(0) \quad i = 1, \dots, M \quad (3.7)$$

where $\bar{Z}'(0)$ (respectively, φ_j^0) are the components of $\psi \bar{Z}(0)$ (respectively, φ^0) in the basis (φ_j) . Let us remark that by construction, the observed state y_0^1 will vanish in $\Omega \setminus \omega$.

B. General Case

In the general case, it is not easy to calculate the eigenfunctions of the operator A . Here we shall give a direct approach which allows to overcome this difficulty and leads to the state to be estimated in ω .

We have seen that the regional observability is equivalent to finding φ^0 such that $\psi(0) = \bar{Z}(0)$ in ω , where $\psi(t)$ and $\bar{Z}(t)$ are solutions of (2.15) and (2.17). So the problem turns up naturally to find φ^0 which is the solution of the following problem

$$\begin{cases} \min \|\psi(0) - \bar{Z}(0)\|_{L^2(\omega)}^2 \\ \varphi^0 \in G \end{cases} \quad (3.8)$$

The resolution of this problem can easily be achieved by the direct minimization algorithm.

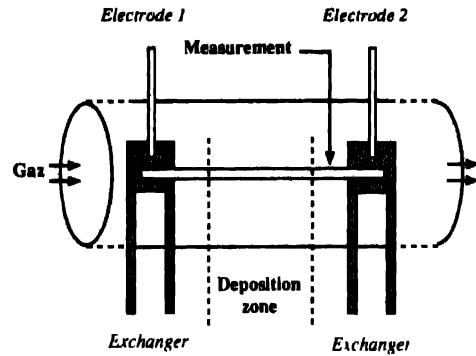


Fig. 1. Longitudinal section of the reactor

Algorithm.

- 1) Solve (2.17) ($\rightarrow \bar{Z}(0)$).
- 2) Initiate φ^0 in G .
- 3) Solve (2.13) ($\rightarrow \varphi(b, t)$).
- 4) Solve (2.15) ($\rightarrow \psi(0)$).
- 5) If $\|\psi(0) - \bar{Z}(0)\|_{L^2(\omega)} > \epsilon$ return to Step 2).
- 6) The observed state on the subregion ω is given by $y_0^1 = \varphi^0$.

Remark The choice of φ^0 can be made by an optimization process which does not need the calculus of the gradient [9]. The initial choice of φ^0 must be as close as possible to the state suggested by the real system. But any arbitrary choice will lead to the same result, with more computations. This is applied to the thermal considered problem.

IV. APPLICATION TO THE THERMAL DEPOSITION PROBLEM

A. System Description

The deposition phenomenon is produced in a reactor with cold walls where a substrate is placed. The temperature is controlled by the energy supplied by electric current (Joule effect) between two electrodes maintained at a constant temperature (they are thermally controlled by a refrigerating fluid circulation). The Joule effect produces a temperature elevation in the substrate which exchanges heat with the surroundings essentially in the radiative form (around a temperature of 1300 degrees C). The deposition concerns only a subregion $\omega \subset \Omega$. See Fig. 1.

The substrate is first steered to the deposition temperature, and then the covering process starts. The surface of the substrate is modified, its emissivity evolves, and then thermal exchanges start. The modification of the thermal profile is to be compensated with action on the electric current. Then it is important to know the variation of this profile around the previous equilibrium state. One of the difficulties is due to the fact that only the temperature out of the deposition zone ω can be measured by optical pyrometry. In view of this, we have to conceive a precise estimator of the temperature profile in the sensitive subregion. This is made by the regional observation techniques developed in the previous sections.

B. Thermal Model

The modelization of the thermal deposition problem can be carried out in one-dimensional space with $\Omega =]0, \ell[$. Let $\omega \subset \Omega$ be the deposition zone. The radiative phenomenon introduces a nonlinear term in the thermal balance which takes the conductive exchange

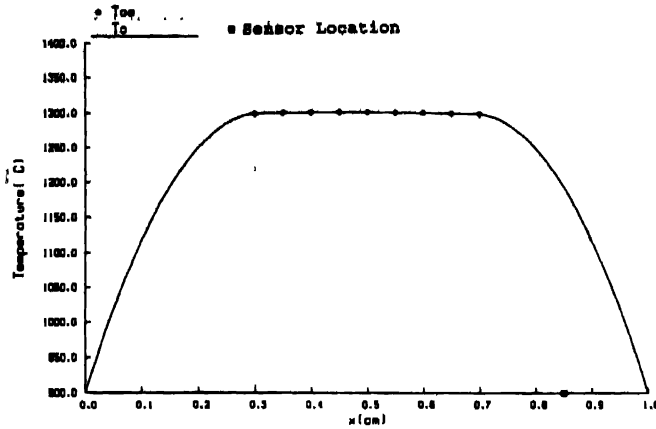


Fig. 2. Comparison of the real state $T_0(x)$ (supposed to be unknown) and the estimated state $T_0(x)$ on the deposition subregion $\omega = [0.3, 0.7]$.

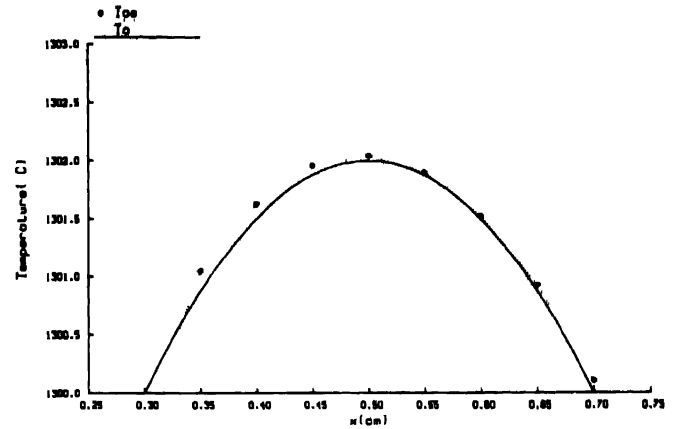


Fig. 3. Zoom of the real and estimated states in the deposition subregion.

into account. This leads to the following modelization of the system

$$\begin{cases} mc_p \frac{\partial T(x, t)}{\partial t} \\ = \lambda \left(c \frac{\partial^2 T(x, t)}{\partial x^2} \right) - 2 \cdot \sigma (t + \epsilon) T^4(x, t) + \frac{\rho}{\epsilon} I^2 &]0, \ell[\times]0, t_f[\\ T(0, t) = T_1 &]0, t_f[\\ T(\ell, t) = T_2 &]0, t_f[\\ T(x, 0) = T_0(x) &]0, \ell[\end{cases} \quad (4.1)$$

where

x : space variable, t : time variable.
 m : mass element by unit length, c_p : calorific heat element.
 λ : thermal conductivity, ρ : electric sensitivity.
 ϵ : emissivity, ℓ : length of the element.
 ϵ : depth of the element, σ : Stephan-Boltzman constant.
 T : temperature of the element, I : current intensity.

C. Linearized Model

The previous model is nonlinear but the nature of the problem allows the study to be made in a linear context. Indeed the deposition is produced around a temperature T_0 . The modification of the thermal profile has to be compensated with a control of the electric current around T_0 . Therefore (4.1) can be linearized around T_0 corresponding to current intensity $I = I_0$.

Let $\Delta T = T - T_0$, $\Delta I = I - I_0$. The linearized model is then

$$\begin{cases} \frac{\partial \Delta T(x, t)}{\partial t} \\ = \alpha \frac{\partial^2 \Delta T(x, t)}{\partial x^2} + \beta T_0^3 \Delta T(x, t) + \gamma &]0, \ell[\times]0, t_f[\\ \Delta T(0, t) = \Delta T(\ell, t) = 0 &]0, t_f[\\ \Delta T(x, 0) = 0 &]0, \ell[\end{cases} \quad (4.2)$$

where α , β , and γ are given by

$$\alpha = \frac{\lambda \ell \epsilon}{mc_p}; \quad \beta = \frac{-8\pi\sigma(t + \epsilon)}{mc_p}; \quad \gamma = \frac{\rho}{mc_p \ell \epsilon} \Delta I \quad (4.3)$$

where ΔI is the control variable and $\lambda, m, c_p, \ell, \epsilon, \sigma, \rho, I_0$ are constants to be specified for numerical simulation.

The problem is to determine the variation of the temperature profile only on the deposition subregion around the previous equilibrium rating. The measurements are given by the output function

$$z(t) = \Delta T(b, t) \quad (4.4)$$

where b is the sensor location, $b \in]0, \ell[\setminus \omega$.

D. Simulation

For numerical simulations we consider the following values for the different parameters

$$\begin{aligned} \Omega &=]0, 1[& \omega &= [\frac{1}{3}, \frac{2}{3}] \\ T_1 &= T_2 = 900^\circ\text{C} & I_0 &= 38 \text{ A} \\ \ell &= 0.510^{-2} \text{ m} & \epsilon &= 10^{-3} \text{ m} \\ z &= 0.2 & c_p &= 61.10^{-3} \text{ J Kg}^{-1} \text{ K}^{-1} \\ \lambda &= 1.03210^2 \text{ W m}^{-1} \text{ K}^{-1} & m &= 1.6 \text{ Kg m}^{-1} \\ \sigma &= 5.7110^{-8} \text{ W m}^{-2} \text{ K}^{-4} & \rho &= 2.8710^{-4} \Omega \text{ m}^{-1}. \end{aligned} \quad (4.5)$$

The sensor is located out of the deposition zone with $b = 0.85$.

Using the algorithm of Section III-B, we obtain the following results: Fig. 2 gives the estimated state (dashed line) and the real state (continuous line).

The regional state in the deposition subregion is estimated with a good accuracy

$$\frac{\|T_0 - T_0\|_{L^2(\omega)}^2}{\|T_0\|_{L^2(\omega)}^2} \simeq 1.410^{-4}. \quad (4.6)$$

REFERENCES

- [1] A. El Jai and M. Amouroux, *Automatique des Systèmes Distribués*. Paris: Hermès, 1990.
- [2] A. El Jai and A. J. Pritchard, *Sensors and Controls in the Analysis of Distributed Systems*. New York: Wiley, 1988.
- [3] A. El Jai, M. Amouroux, and E. Zerrik, "Regional observability of distributed systems," *Int. J. Sc. Sci.*, vol. 25, no. 2, pp. 301-313, 1994.
- [4] A. El Jai, M. C. Simon, and E. Zerrik, "Regional observability of distributed systems," *Int. J. Sensor and Actuators*, vol. 39, no. 2, pp. 95-102, 1993.
- [5] J. L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués*. Paris: Masson, 1988.
- [6] A. B. Devrient, *La Transmission de la Chaleur*, Gaëtan Morin, Ed. Paris: 1984.
- [7] D. W. Hess, K. F. Jensen, and T. J. Anderson, "Chemical vapor deposition: a chemical engineering perspective," *Rev. Chem. Eng.*, vol. 3, pp. 97-186, 1985.
- [8] N. Reffe, "CVD: Validation de modèle et asservissement de température," *Rapport IMP*, Perpignan, Sep. 90.
- [9] M. J. D. Powell, *Computer J.*, vol. 7, no. 2, 1964.

Towards a Generalized Regulation Scheme for Oscillatory Systems via Coupling Effects

Kevin L. Tuer, M. Farid Golnaraghi, and David Wang

Abstract—The vibration suppression laws presented in this work utilize an energy transfer phenomena that is evident in some linearly and nonlinearly coupled systems. The first iteration of the formulation is presented in this paper. These control laws are unique in that the states approach the operating point on the order of a cosine function, the nonlinear control law renders a unidirectional control input, and the intuition associated with the energy transfer analogy aids the controller design.

I. INTRODUCTION

Many techniques have been formulated to suppress oscillations in both linear and nonlinear systems. The standard linear time-invariant state-space representation of a system has the form

$$\begin{aligned}\dot{x} &= Ax + Bu \\ y &= Cx + Du\end{aligned}\quad (1)$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$ and $D \in \mathbb{R}^{p \times m}$. In contrast, a typical nonlinear system is represented by

$$\begin{aligned}\dot{x} &= f(x, u) \\ y &= g(x, u)\end{aligned}\quad (2)$$

where $f: \mathbb{R}^{n+m} \rightarrow \mathbb{R}^n$ and $g: \mathbb{R}^{n+m} \rightarrow \mathbb{R}^p$. The corresponding control techniques can also be classified as linear or nonlinear. Included in the linear realm are techniques such as the standard PID control [1], pole-placement control [1], and linear quadratic (LQ) techniques [1]. The latter involves the minimization of a quadratic cost function to define the control law. More recently, the H_∞ optimal control technique [2] has been used quite extensively. It entails the selection of a controller from the set of all possible stabilizing controllers by minimizing the infinity operator norm of the transfer function between the exogenous inputs and exogenous outputs of a standard augmentation. Structurally similar approaches include the H_2 [2] and L_1 [3] control methods. The field of nonlinear control includes techniques such as the Lyapunov approach [4], the passivity formalism [4], and dissipative system theory [5]. These particular strategies have energy-based scenarios. The Lyapunov approach involves specifying a control input such that the candidate Lyapunov function is locally positive definite and its time derivative locally negative (semi) definite. A similar approach is taken using the passivity formalism which involves the use of a Lyapunov-like function that is an indicator of the energy content of the system at any point in time and space. The control input to the system is selected such that the system is passive and/or dissipative. In addition, closed-loop stability of a feedback system can be ascertained using the well known passivity theorem [4]. Dissipative system theory, in contrast, employs energy storage and supply functions to model the energy

based characteristics of a system. The control input to the plant is chosen to ensure that the closed-loop system is dissipative which translates to a decrease in the system energy. Other nonlinear control techniques involve the use of a variable feedback law including sliding mode [4] and gain scheduling [6]. For example, the former involves specifying a sliding surface in the state space and defining the appropriate control parameters such that a sliding mode exists and the surface is locally or globally attractive. Implementation typically involves high speed switching feedback which can induce spillover effects. Gain scheduling is similar except the transition between feedback structures is smoother [6]. Some nonlinear systems support the use of feedback linearization techniques [4]. Using this approach, the designer develops a feedback law that essentially linearizes the plant. Then, a wealth of linear control algorithms can be utilized to complete the control design. These techniques, however, tend to suffer from a lack of robustness in the presence of uncertainty. More recently, neural networks [7] and fuzzy logic [7] have been utilized for vibration suppression applications as well.

Implementation of the control techniques outlined above may render a system that is stable in an L_2 , L_∞ , or exponential sense. Stability in this sense, however, does not preclude the possibility of one or more of the states of the controlled system exhibiting an unacceptable rate of convergence to the desired state and/or unwanted oscillatory behavior. The techniques presented in this work are intended to reduce or eliminate these effects.

For the purposes of this paper, the investigation is conducted using a canonical second order system with a state-space model given by

$$\dot{x}_p = \begin{bmatrix} 0 & 1 \\ -\omega_p^2 & -2\zeta_p\omega_p \end{bmatrix} x_p + f_p(x_p) + \begin{bmatrix} 0 \\ b_{21} \end{bmatrix} u \quad (3)$$

where $\zeta_p, \omega_p \in \mathbb{R}^+$ are the damping coefficient and natural frequency respectively, $b_{21} \in \mathbb{R}$, $f_p(x)$ contains weak nonlinear terms and the state vector is $\{x_p\} = [x_1 \ x_2] = [x \ \dot{x}]$. Equation (3) can be considered as the model of one oscillatory mode of a multi-degree-of-freedom (MDOF) system. Thus, the intention is to develop the theory for a canonical oscillatory system with the ultimate goal of utilizing the theory for multi-mode oscillation suppression. For notational purposes, the system given in (3) will be denoted as the plant.

The essence of the proposed techniques is to exploit the energy link that can exist between two coupled, second order, oscillating systems. If one equation models a system to be controlled and the other models a controller, then it is plausible to alter the dynamic characteristics of the plant using the controller. Under this scenario, the plant and the controller dynamics are characterized by amplitude and/or phase modulation giving rise to a beat phenomenon. As a result, the envelope of the plant response is periodic and approaches a minimum at approximately the same rate as a cosine function. Subsequent controller action can then be instituted to "remove" energy from the entire system. Using the energy transfer analogy, the design is well directed and efficient. The designer is able to incorporate physical meaning with the mathematics to yield a control strategy that is intuitive and relatively easy to troubleshoot.

The use of coupling effects to suppress structural oscillations has been addressed recently in [8]–[14]. Golnaraghi was one of the first researchers to use modal coupling effects to control structural oscillations [8], [9]. Tuer *et al.* [10], Duquette *et al.* [11], Duquette [12], and Tuer [13] examined the use of coupling effects to control

Manuscript received March 12, 1993; revised April 20, 1994. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

The authors are with the ConStruct Group University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

IEEE Log Number 9407230.

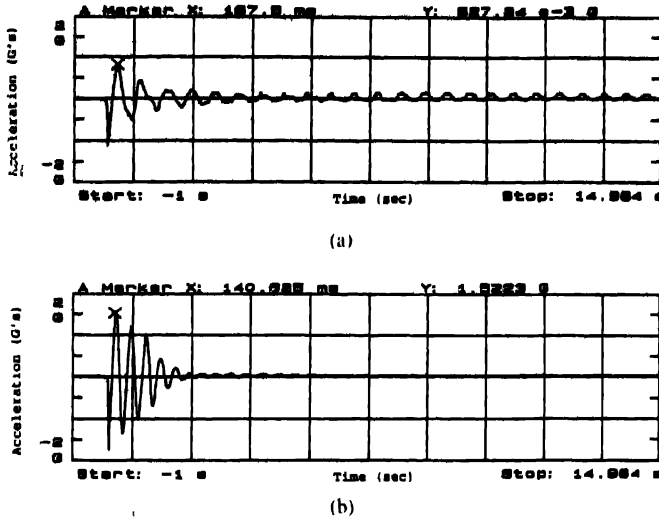


Fig. 1. Experimental Responses [12]. (a) Coordinate coupling, (b) modal coupling.

oscillations in a flexible cantilever beam on both a theoretical and experimental level. The desired coupling effects were achieved via physical components. A DC motor was attached to the free end of the beam. A rigid beam-mass assembly was connected directly to the shaft of the motor. This assembly was designated the secondary link. The position of the secondary link was regulated at 90 degrees to the flexible beam via position and velocity feedback from the motor. Fig. 1(a) shows the experimental plot of the regulated response of the tip of the flexible beam using the coordinate coupling control law after subjecting the tip to an initial displacement. Similarly, Fig. 1(b) illustrates the tip response employing the modal coupling control law. Both strategies function very effectively. These experimental results illustrate that physical systems can exhibit both coordinate and modal coupling effects under the proper conditions. This investigation also showed that these phenomena can be used, rather effectively, for control purposes. This physical insight provides the primary motivation for the current research.

The disadvantage of the aforementioned work is the need for unconventional actuation devices to apply the necessary control. This problem has been addressed by Tuer *et al.* [14]. They have initiated the generalization of the use of modal coupling effects to control a canonical system using more conventional actuation devices. In this paper, both linear and nonlinear coupling effects are employed as paradigms in the development of control laws to regulate an oscillatory state of a system. As well, stability issues for both control strategies are addressed. A numerical example is presented to illustrate the use of the proposed techniques and corroborate the mathematical predictions.

II. CONTROLLER IMPLEMENTATION

To suppress the oscillations of a system of the form of (3) using the proposed techniques, a second system must be introduced to which the plant can be coupled using feedback. This is accomplished via the introduction of second-order differential equation with a linear structure similar to that of the plant. This system becomes the controller in the closed-loop system. Thus, the closed-loop system dynamics are dictated by equations of the form

$$\dot{x}_p = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_p^2 & -2\zeta_p\omega_p \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + f_p(x_p) + \begin{bmatrix} 0 \\ b_{21} \end{bmatrix} u \quad (4)$$

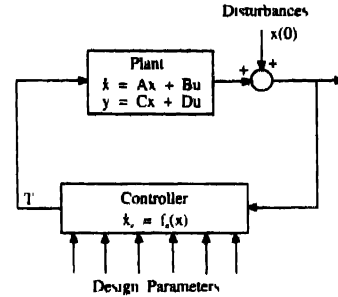


Fig. 2. Block diagram of controlled system.

$$\dot{x}_c = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_c^2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_1 \end{bmatrix} + \begin{bmatrix} 0 \\ C'(x_2, x_1) \end{bmatrix} + f_c(x_p, x_c) \quad (5)$$

where (4) models the plant, (5) models the controller and $\omega_c \in \mathbb{R}^+$ is the controller frequency. The function $C'(\cdot)$ is a damping mechanism utilized to effect control and $f_p(\cdot)$ contains the plant nonlinearities. The plant control input u , which takes the form of a state relation in the formulation, and $f_c(\cdot)$ are used to establish the coupling between the plant and the controller. The coefficient of the relation for u , the elements of $f_c(\cdot)$, the form of $C'(\cdot)$, and the controller frequency ω_c are the design parameters of the system. A block diagram of the closed-loop system is illustrated in Fig. 2. As evident from the figure, the disturbance is modelled by initial state conditions acting at the output of the plant.

The design of the controller is a two-stage process. The first step involves establishing the required coupling effects. The second step revolves around specifying the remaining design parameters to establish the frequency of amplitude modulation and defining the appropriate parameters of the energy removal mechanism such that the disturbance induced motion is quelled in accordance with the given design specifications.

A. Establishing Coupling Effects

The proposed technology hinges on both linear and nonlinear coupling effects to function. In the past, the phenomena resulting from coupling effects were examined primarily from an analysis standpoint. However, the focus of this work is to utilize the desirable features of these phenomena for control purposes.

1) *Coordinate Coupling*: In a mathematical sense, coordinate coupling refers to the nondiagonal nature of the inertia and/or stiffness matrices of a system of second order, ordinary differential equations of motion. In certain situations, the system response may exhibit periodic amplitude modulation or a beat phenomenon. Since the solutions of (4) and (5) are predominantly oscillatory, this amplitude modulated response must be the result of the superposition of "tuned" periodic functions. This can be illustrated using the analogy of the pointwise addition of two periodic functions with the same amplitude and nearly the same frequencies of oscillation. There are well-defined points in time where the two functions reinforce each other to yield a doubling of the amplitude. Conversely, there are points in time where the two functions act to cancel one another yielding a net amplitude of zero. On a physical level, this is analogous to a transfer of energy between the coordinates of the system.

The development of the control law based on coordinate coupling was conducted by first recognizing that a similar phenomena can be established between a second order, single-degree-of-freedom (SDOF) plant and second order, SDOF controller through the proper selection of the controller parameters and the control input to the

plant. That is, the control law was established so as to generate a closed-loop system with inertial coupling. To this end, the closed-loop equations have the form of (4) and (5) with

$$\begin{aligned} u &= -\dot{x}_1 \\ f_r(\cdot) &= -K\dot{x}_2 \\ f_p(\cdot) &= 0 \end{aligned} \quad (6)$$

where $K \in \mathbb{R}^+$. In general, the last condition in (6) need not be satisfied. It is implicitly assumed, however, that any nonlinearities contained in $f_p(\cdot)$ are at least one order of magnitude smaller than the linear terms of the closed-loop system equations. The solutions of (4) and (5) with (6) invoked, assuming zero damping, are readily found to have the form

$$\begin{bmatrix} x_1 \\ x_3 \end{bmatrix} = \begin{bmatrix} C_1 A_1^1 & C_2 A_1^2 \\ C_1 A_2^1 & C_2 A_2^2 \end{bmatrix} \begin{bmatrix} \cos(\omega_1 t + \Phi_1) \\ \cos(\omega_2 t + \Phi_2) \end{bmatrix} \quad (7)$$

where $A_j^i \in \mathbb{R}$ is the i th entry of the j th eigenvector corresponding to the j th natural frequency of the closed-loop system, $\omega_j \in \mathbb{R}^+$. The constants $C_j \in \mathbb{R}$ and $\Phi_j \in \mathbb{R}$ are defined by the initial conditions imposed on the system. The eigenvalues of the undamped system $\lambda_{1,2} \in \mathbb{C}$, are

$$\lambda_{1,2} = \frac{(\omega_p^2 + \omega_c^2) \pm \sqrt{-(\omega_p^2 + \omega_c^2)^2 - 4(1-K)(\omega_p^2 \omega_c^2)}}{2(1-K)} \quad (8)$$

The degree of coupling and frequency of amplitude modulation is contingent on the magnitude of K since this parameter dictates the proximity of the system's natural frequencies.

The first solution in (7) models the plant response whereas the second solution pertains to the controller response. Using the design parameter, K , the natural frequencies of the closed-loop system can be placed arbitrarily close to one another. Since the plant and controller responses are the result of the superposition of two periodic functions with nearly the same frequencies, a linear beat phenomenon can be established. Under these conditions, the periodic functions will completely cancel one another provided the amplitudes of the two functions are equivalent. In terms of the solution for the plant response, this translates to enforcing the condition $C_1 A_1^1 = C_2 A_1^2$ in (7). In [12], it was shown that this condition is satisfied if

$$\omega_c = \omega_p. \quad (9)$$

To ensure a relative plant energy minimum at a specific point in time, the parameter K is chosen so as to satisfy the relation [12]

$$n = \frac{\omega_2}{\omega_2 - \omega_1} \quad (10)$$

where the user-defined $n \in \mathbb{R} \setminus 0$ dictates the amplitude modulation frequency. Both (9) and (10) are necessary conditions to establish the desired amplitude modulated response. The time at which the first cancellation occurs, t_{fc} , can be shown to occur at [12]

$$t_{fc} = \frac{\pi}{\omega_2 - \omega_1}. \quad (11)$$

Thus, satisfying (9) ensures that $C_1 A_1^1 = C_2 A_1^2$ to foster complete cancellation at time t_{fc} , the value of which is dictated, primarily, by the magnitude of n as expressed by (10). That is, the parameter n is used to define the point in time where the position and velocity of the plant are simultaneously zero which corresponds to zero plant energy.

In the foregoing formulation, acceleration feedback was utilized to establish dynamic coupling effects. Analysis shows that a static coupling paradigm can also be utilized and that both (9) and (10) hold under this scenario as well [15], [16]. Since both types of coupling render the same characteristics, the remainder of the analysis is conducted for the dynamic coupling scenario only.

2) Modal Coupling: A nonlinear system may exhibit modal coupling effects if it is in a state of internal resonance. Internal resonance has its foundation in the field of nonlinear oscillations. Mathematically, a necessary but not sufficient condition for the existence of internal resonance is defined by the commensurability of the linear natural frequencies of the nonlinear equations of motion, ω , [17]. That is, there exist constants $\{\Psi_1, \dots, \Psi_n \in \mathbb{R} \setminus 0\}$ such that

$$\Psi_1 \omega_1 + \Psi_2 \omega_2 + \dots + \Psi_n \omega_n \approx 0. \quad (12)$$

Under such conditions, an energy bridge is formed between the commensurable modes of oscillation. If the left-hand side and right-hand side of (12) are identically equal, an exact commensurability condition exists which fosters a state of maximum energy transfer.

The type of internal resonance condition is contingent on the order of nonlinearities. For instance, if quadratic nonlinearities are present in the system, a 2:1 internal resonance condition may exist. The specific commensurability relation under this scenario is $\Psi_1 \omega_1 + \Psi_2 \omega_2 \approx 0$ where the associated constants are $\{\Psi_1 = \pm 2, \Psi_2 = \mp 1\}$ or $\{\Psi_1 = \pm 1, \Psi_2 = \mp 2\}$. A state of 2:1 internal resonance is employed in this paper.

As with the coordinate coupling-based control law presented in the previous section, the internal resonance-based control law was developed by recognising that a state of internal resonance can be established between a second order, SDOF plant and a second order, SDOF controller. Thus, using this paradigm, the closed-loop system equations are of the form given by equations (4) and (5) with

$$\begin{aligned} u &= -\gamma_1 x_1^2 \\ f_r(\cdot) &= \gamma_2 x_2 x_1 \\ f_p(\cdot) &= 0 \\ \omega_c &\approx \frac{1}{2} \omega_p \end{aligned} \quad (13)$$

where $\gamma_1, \gamma_2 \in \mathbb{R}^+$ are design parameters. In general, $f_p(\cdot)$ need not be zero. In this case, if the structure of the nonlinear terms, $f_p(\cdot)$, parallel the nonlinearity introduced by the control law, they can be used to effect the control. If these terms have a structure differing from that of the nonlinearities introduced by the control, they will not affect the integrity of the controller provided they are weak relative to the control law nonlinearities. Finally, the last condition in (13) is employed to satisfy the commensurability requirements and is necessary for establishing modal coupling effects.

Other forms of nonlinear feedback can be utilized in the formulation of the control law [15]. That is, the nonlinear terms remain quadratic but the individual terms may include position, velocity, and/or acceleration feedback. The analysis and resulting relations differ slightly but, nevertheless, have essentially the same structure. Therefore, as with the linear case, the potential to use other forms of feedback renders more freedom to the design. In this paper, the system given by (4), (5), and (13) is analyzed and discussed in detail.

To illustrate the modal transfer of energy and to derive a relation indicating the degree of energy transfer, a perturbation analysis technique known as the method of multiple scales is utilized [17]. Using this approach, an assumed solution for the undamped nonlinear system is constructed in the form of an asymptotic series. Several relations indicating the existence of internal resonance are derived in the process of formulating the exact form of the elements of the series solution. The system solution is of the form

$$\begin{aligned} x_1 &= x_{10}(T_0, T_1) + \epsilon x_{11}(T_0, T_1) + \dots \\ x_2 &= x_{20}(T_0, T_1) + \epsilon x_{21}(T_0, T_1) + \dots \end{aligned} \quad (14)$$

where T_i are new independent time scales and the corresponding asymptotic sequence is constructed using powers of $\{\epsilon \in \mathbb{R}^+ \mid 0$

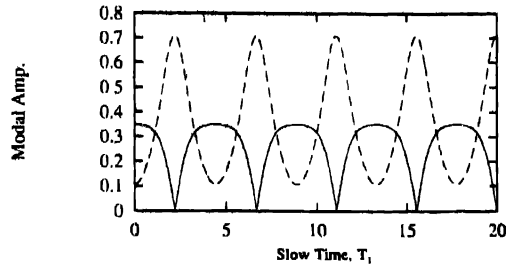


Fig. 3. Undamped Modal Response. — : R_1 , - - - : R_2

$\epsilon \ll 1$. Substituting (14) into the undamped form of the nonlinear equations of motion yields a set of conditions that must be satisfied in order to satisfy the constraints implicit on the series solution. These are known as the secular terms or solvability conditions and can be expressed in terms of the modal amplitudes, R_1 , R_2 , and associated phases, α_1 , α_2 , of the system

$$\begin{aligned}\dot{R}_1 &= -0.0625\gamma_1\omega_p R_2^2 \sin(\phi) \\ \dot{R}_2 &= 0.25\gamma_2\omega_p R_1 R_2 \sin(\phi) \\ R_1 \dot{\phi} &= (0.5\gamma_2\omega_p R_1^2 - 0.0625\gamma_1\omega_p R_2^2) \cos(\phi) \\ \phi &= 2\alpha_2 - \alpha_1\end{aligned}\quad (15)$$

where the overdot () indicates differentiation with respect to the time variable T_1 . The variables R_1 and R_2 model the envelope of the response of the plant and controller, respectively. Equations (15) model the response of the envelopes of the modes of oscillation of the system. Simulation of these equations indicate that the desired transfer of energy between modes is indeed occurring as illustrated in Fig. 3. The corresponding parameter values are $\{\omega_p = 30, \gamma_1 = \gamma_2 = 0.5\}$. Equations (15) are used to generate an energy relation which has the form

$$1 \frac{\gamma_2}{\gamma_1} R_1^2 + R_2^2 = E \quad (16)$$

where $E \in \mathbb{R}^+$ is representative of the system energy and R_i^2 is indicative of the energy of mode i . Equation (16) shows that, provided R_1 is a periodic function, a continuous exchange of energy occurs in this conservative system. The secular term equations are further analyzed to generate a relation used as an indicator of the degree of energy transfer within the system. This relation takes the form

$$K_1 = \frac{4\gamma_2 [R_2(0)^2 R_1(0) \cos(\phi(0))]^2}{\gamma_1 [4 \frac{\gamma_2}{\gamma_1} R_1(0)^2 + R_2(0)^2]} \quad (17)$$

where $K_1 \in \mathbb{R}^+$. In general, the smaller the value of K_1 , the better the transfer of energy between modes. Perturbation analysis shows that $K_1 = 0$ corresponds to an unstable equilibrium point and is therefore inadmissible. It appears after extensive numerical analysis that (17) can be extrapolated to the full-order nonlinear system

$$\frac{4\gamma_2 [x_1(0)^2 x_3(0) \cos(\phi(0))]^2}{\gamma_1 [4 \frac{\gamma_2}{\gamma_1} x_1(0)^2 + x_3(0)^2]} \leq \frac{4\gamma_2 [x_3(0)^2 x_1(0)]^2}{\gamma_1 [4 \frac{\gamma_2}{\gamma_1} x_1(0)^2 + x_3(0)^2]} = K_1. \quad (18)$$

Thus, (18) illustrates the effect that the disturbance and the design parameters have on the degree of energy transfer. This is an important result because K_1 and (18) can be utilized in the formulation of the controller. To guarantee a particular level of energy transfer, K_1 is selected arbitrarily small and the design parameters are chosen to satisfy (18). In addition, actuator constraints can be incorporated into the design at this stage since a smaller value of K_1 translates to larger control effort. Further detail of the perturbation analysis and the derivation of (17) is contained in [14].

B. Achieving Vibration Suppression

To effect control over the system, the energy must be transferred from the plant to the controller and "removed" before it has an opportunity to transfer from the controller back to the plant. To accomplish this task, two fundamentally different approaches are proposed [14]. The first method is the disabled input method (DIM). Using this method, the existing energy bridge is severed once the energy has transferred from the plant by disabling the control signal to the plant. The second method is termed the dissipated energy method (DEM). This approach, in essence, dissipates the energy as it is transferred from the plant to the controller via a damping mechanism associated with the controller.

1) *Disabled Input Method (DIM)*: An outline of the design procedures for the coordinate and modal coupling controllers are delineated separately beginning with the former.

a) *Coordinate coupling controller*: Using this paradigm of control, the governing equations of the closed-loop system are given by (4)–(6) with $C(\cdot) \equiv 0$. The first step of the procedure is to set the controller frequency, ω_c , equal to the natural frequency of the plant, ω_p , to facilitate a maximum energy transfer. The desired response frequencies are established using the controller gain K since the natural frequencies of the closed-loop system are explicitly a function of K . The frequency of amplitude modulation and primary response frequency are

$$\left(\frac{\omega_2 - \omega_1}{2} \right), \left(\frac{\omega_2 + \omega_1}{2} \right) \quad (19)$$

respectively, where the former is an indication of the settling time of the system.

Since the proposed techniques hinge on energy transfer, it is necessary to construct an energy function for the plant to effect control. In general, the energy function has a quadratic form and is given by

$$\mathbf{x}_p^T Q \mathbf{x}_p \quad (20)$$

where Q is a real, diagonal, positive definite weighting matrix

$$Q = \begin{bmatrix} \frac{\omega_p^2}{2} & 0 \\ 0 & 1 \end{bmatrix}. \quad (21)$$

The final task of the design involves monitoring the plant's energy on a continuous basis and disabling the input to the plant when the magnitude of the energy function falls below a predetermined threshold value, $E_i^{DIM} \in \mathbb{R}^+$. That is

$$u = \begin{cases} -i_1, & \mathbf{x}_p^T Q \mathbf{x}_p > E_i^{DIM} \\ 0, & \mathbf{x}_p^T Q \mathbf{x}_p \leq E_i^{DIM} \end{cases} \quad (22)$$

b) *Modal coupling controller*: Under this scenario, the closed-loop dynamics are dictated by (4), (5), and (13). The procedure resembles the latter with a few fundamental differences. First of all, the frequency of the controller, ω_c , is set equal to one-half the natural frequency of the plant, ω_p . Next, with $(\frac{\gamma_2}{\gamma_1}) = 1$ and an arbitrarily small k , the required controller initial condition, $x_3(0)$, is calculated using (18). To attain the desired amplitude modulation frequency, maintain $(\frac{\gamma_2}{\gamma_1}) = 1$ and set γ_1, γ_2 such that the point at which the amplitude envelope achieves a minimum value corresponds to the desired settling time. At this point, double peak symmetry in the plant response is achieved to maximize the dwell time at the plant energy minimum thus rendering the system more robust to errors in disabling time. This is done by adjusting the parameter $x_3(0)$. Using this new value of $x_3(0)$ and the specified design parameters, the degree of modal energy transfer, K_1 , is recalculated using (18). If the energy transfer corresponding to this value of K_1 is insufficient, the value of K_1 is reduced and the previous steps are repeated. Otherwise,

the final stage of the design entails monitoring the system energy and disabling the input in a similar manner described in the previous section. That is

$$u = \begin{cases} -\gamma_1 x_4^2, & x_p^T Q x_p > E_t^{DIM} \\ 0, & x_p^T Q x_p \leq E_t^{DIM} \end{cases} \quad (23)$$

Although effective, this technique suffers from lack of robustness to errors in disabling time. When the input is disabled, the energy bridge between the plant and the controller is destroyed. Thus, whatever energy is in the plant mode when the input is disabled remains in the plant mode unless controller action is reinitiated.

b) Dissipated energy method (DEM): The design procedure for the DEM is similar to the DIM except for the means of oscillation suppression. Using this approach, a damping model is incorporated into the controller equation. The energy link facilitates the transfer of energy from the plant to the controller where it is dissipated through the controller's damping mechanism. This technique can be implemented in either a time invariant fashion or a time-varying fashion. The time invariant approach involves invoking the damping mechanism associated with the controller in conjunction with the disturbance. That is, given that the disturbance is applied at time t_d and the damping mechanism is invoked at time t_{imp} , then $t_{imp} = t_d$. Thus, the system approaches the operating point slower than a cosine function but quicker than an exponential function. The time-varying approach involves the use of a function that provides an on-line indication of the plant energy. Once this indicator function drops below a user-specified threshold, the damping effects are invoked. Thus, given that t_{min} denotes the time at which the plant reaches its first energy minimum, then $\{t_{imp} = t_{min} - d : 0 < d \leq (t_{min} - t_d), d \in \mathbb{R}^+\}$. Alternatively, the damping invoke condition can be energy based. That is

$$C(\cdot) = \begin{cases} 0, & x_p^T Q x_p > E_t^{DEM} \\ C_{d,s}(\cdot), & x_p^T Q x_p \leq E_t^{DEM} \end{cases} \quad (24)$$

where E_t^{DEM} is the user-specified energy threshold and $C_{d,s}(\cdot)$ is the tuned damping mechanism rendering the desired response. This implementation allows the envelope of the response to approach the operating point with the speed of a cosine function until the switching condition is invoked at which point the damping effects are implemented. From this time forward, the speed of approach becomes similar to that of the time invariant approach. The threshold, $E_t^{DEM} \in \mathbb{R}^+$, must be set at a value that guarantees that the damping effects are employed. That is, if the plant energy does not reach the threshold value, the damping effects will not be invoked and the desired response will not be achieved.

Unlike the DIM, the DEM technique is relatively robust. More details of the design procedure can be found in [14].

C. Stability Analysis

The stability of the closed-loop system can be ascertained for both the coordinate coupling and modal coupling scenarios. For the DEM, a viscous damping mechanism (i.e., $C(\cdot) = 2\zeta_c \omega_c x_4$) is implemented in the controller equation.

1) Coordinate Coupling Controller: System stability using the linear feedback approach is attained by examining the eigenvalues of the closed-loop system which, in the absence of damping ($\zeta_p = 0, \zeta_c = 0$) are dictated by (8). Thus, the undamped closed-loop system is globally stable provided $0 < K < 1$. If the time invariant DEM approach is utilized with $\zeta_p > 0, \zeta_c > 0$, it can be shown that the system is globally asymptotically stable provided $0 < K < 1$. Stability for the DIM and the time-varying DEM is more difficult to show. Since, however, the closed-loop system is stable when the control is applied and after the control is disabled using the DIM, it is

plausible to conjecture stability based on intuition. A similar stability conjecture can be made for the time-varying DEM since the closed-loop system is stable before and after the damping mechanism is invoked.

2) Modal Coupling Controller: The fundamental approach to ascertaining the stability of a nonlinear system is to use Lyapunov's indirect method [4]. In this case, the closed-loop system equations are linearized about the origin of the state-space. If the time invariant DEM is utilized and thus $\zeta_p > 0, \zeta_c > 0$, the system is asymptotically stable. If, however, $\zeta_c = 0$ and/or $\zeta_p = 0$, then the stability of the system cannot be ascertained since at least one of the eigenvalues acquires a zero real part. Thus, Lyapunov's direct method is adopted [4]. The existence of quadratic nonlinearities in the closed-loop system suggests the use of a candidate Lyapunov function of the form

$$V(x) = ax_1^2 + bx_2^2 + cx_3^2 + dx_4^2 \quad (25)$$

where $a, b, c, d \in \mathbb{R}^+$. Taking the time derivative of (25) and substituting the expressions for the derivatives of the state variables yields

$$\begin{aligned} \dot{V}(x) = & -4\zeta_p \omega_p x_2^2 - 4d\zeta_c \omega_c x_1^2 + (2a - 2b\omega_p^2)x_1 x_2 \\ & + (2d\gamma_2 - 2b\gamma_1)x_2 x_4^2 + (2c - 2d\omega_c^2)x_3 x_4. \end{aligned} \quad (26)$$

Selecting the variables of the candidate function to be

$$a = \omega_p^2, \quad b = 1, \quad c = \left(\frac{\gamma_1}{\gamma_2}\right)\omega_c^2, \quad d = \left(\frac{\gamma_1}{\gamma_2}\right) \quad (27)$$

induces the derivative of the Lyapunov function to take the form

$$\dot{V}(x) = -4\zeta_p \omega_p x_2^2 - 4\left(\frac{\gamma_1}{\gamma_2}\right)\zeta_c \omega_c x_1^2 \quad (28)$$

which shows that the closed-loop system is stable over the entire state-space provided

$$\zeta_p, \zeta_c \geq 0, \quad \omega_p, \omega_c, \left(\frac{\gamma_1}{\gamma_2}\right) > 0. \quad (29)$$

As with coordinate coupling, stability for the modal coupling DIM and the time-varying DEM is difficult to prove. Closed-loop stability can be conjectured based on a similar argument as given in the previous section.

Asymptotic stability for the nonlinear feedback configuration cannot be ascertained due to the fact that once the plant is motionless or the equation associated with the controller becomes inactive, the modal energy bridge is destroyed. The energy bridge can be regenerated at any time, however, provided both the plant and controller are active. Thus, it may be possible to prove asymptotic stability if a small, persistently exciting signal is applied to the controller equation for the duration of plant activity.

III. NUMERICAL EXAMPLES

In this section, the proposed control techniques are applied to a plant of the form

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ -900 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \quad (30)$$

where $\zeta_p, f_p(\cdot) = 0$ have been assumed for convenience. The uncontrolled plant response is given in Fig. 4. If coordinate coupling effects are utilized, the control structure is

$$u = -\dot{x}_4, \quad f_c(\cdot) = -K\dot{x}_2, \quad \omega_c = \omega_p. \quad (31)$$

The plant and controller responses when (31) are enforced and $h = 0.0017$ are given in Fig. 5. Clearly, the desired transfer of energy is occurring. If the value of K is increased to $K = 0.005$, which corresponds to $n = 425$, the first energy minimum of the plant occurs

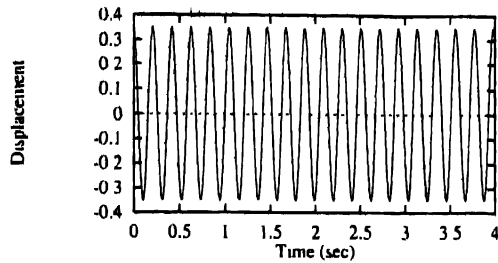
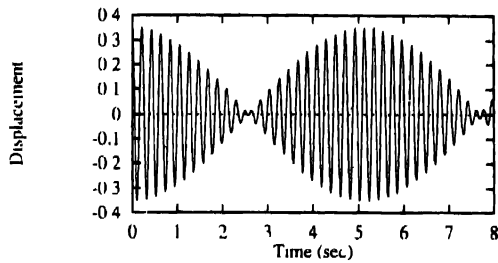
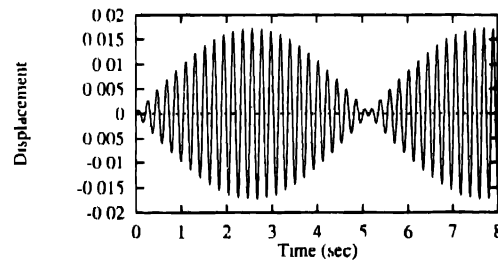


Fig. 4. Uncontrolled plant response.



(a)



(b)

Fig. 5 System response using coordinate coupling (a) Plant, (b) controller.

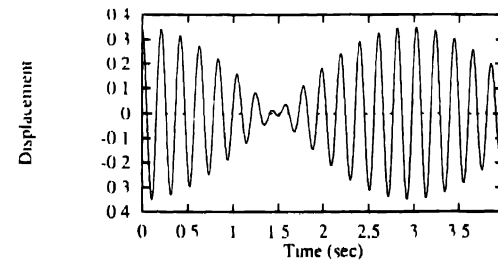


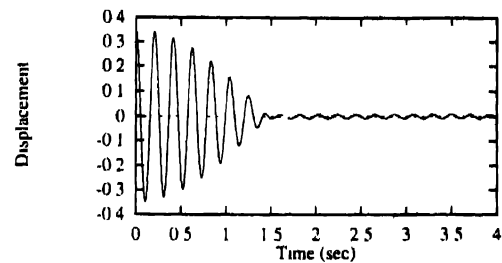
Fig. 6 Plant response, coordinate coupling, with desired amplitude modulation frequency

at approximately 1.5 seconds as illustrated in Fig. 6. Utilizing the DIM results in the controlled response and input profile illustrated in Fig. 7. The steady-state oscillation is an indication of the amount of energy in the plant when the input was disabled. If the DEM approach is adopted and a viscous damping mechanism is utilized, a time invariant implementation results in the plant response and input profile given in Fig. 8 whereas if the time-varying approach is utilized, the plant response and input profile are as given in Fig. 9.

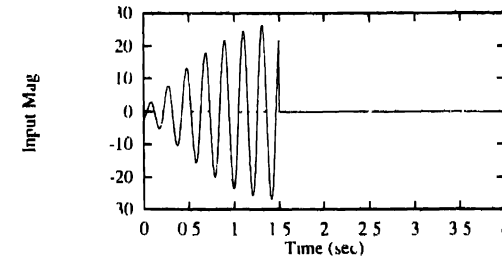
If modal coupling effects are employed, the control parameters are

$$u = -\gamma_1 \dot{x}_1^2, f_1(\cdot) = \gamma_2 x_2^2 x_1, \omega_c = \frac{1}{2} \omega_p. \quad (32)$$

The corresponding plant and controller response with (32) invoked is illustrated in Fig. 10. The magnitudes of the nonlinear coefficients are then increased to render a plant response with the first energy minimum occurring at approximately 1.5 seconds (Fig. 11). To achieve symmetry in the response, the initial position condition

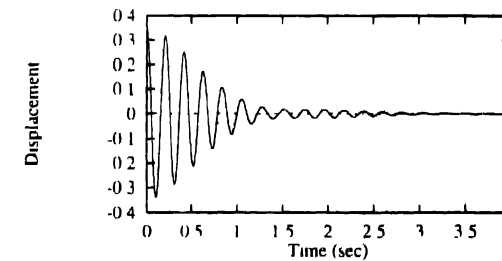


(a)

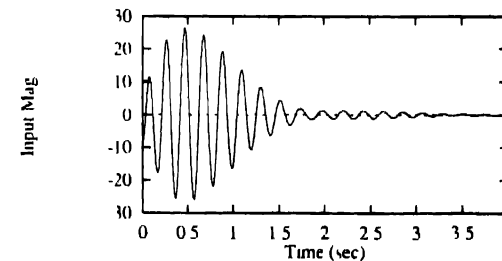


(b)

Fig. 7 DIM controller using coordinate coupling. (a) Plant response, (b) input.



(a)



(b)

Fig. 8 DEM time invariant controller using coordinate coupling. (a) Plant response, (b) input.

is increased. The resulting response is illustrated in Fig. 12. Finally, implementing the DIM control law yields the plant response and input profile given in Fig. 13. With the given choice of nonlinear feedback, the control input is always on one side of zero. This may be advantageous for some applications. If the plant is controlled using the time invariant DEM, the plant response and input profile are as given in Fig. 14. If the time-varying approach is implemented, however, the plant response and control input profile are as given in Fig. 15. Fig. 16 illustrates the controlled response and input profile of the plant subject to a proportional-derivative control law. These are included for comparison purposes. A complete list of parameters for the coordinate coupling and modal coupling controllers is contained in Tables I-III.

In all cases, the proposed control strategies function very effectively. Unlike many techniques, the controller design is aided by the

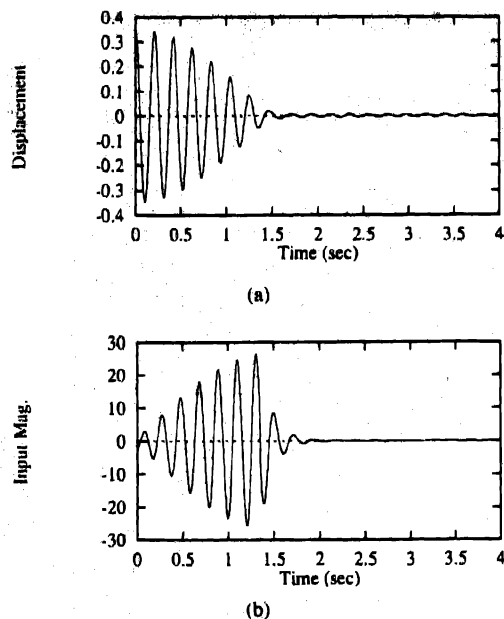


Fig. 9. DEM time-varying controller using coordinate coupling. (a) Plant response, (b) input.

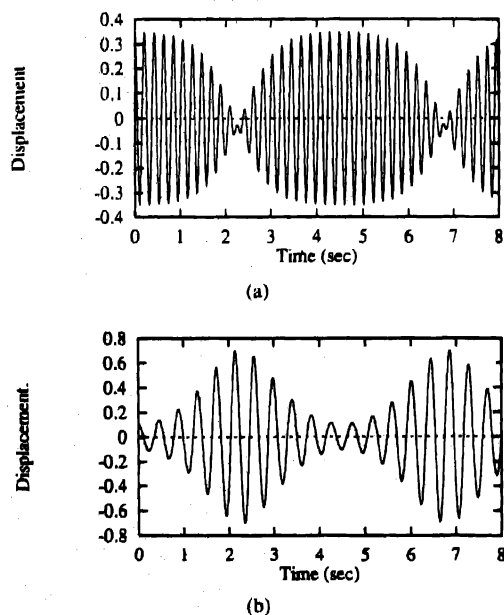


Fig. 10. System response using modal coupling. (a) Plant, (b) controller.

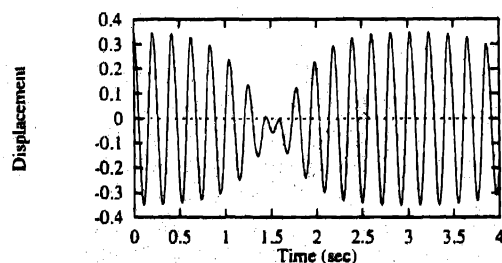


Fig. 11. Plant response, modal coupling, with desired amplitude modulation frequency.

intuition accompanying the energy transfer analogy upon which the derivation of the control laws were based. Thus, troubleshooting and fine tuning tasks are relatively easy to perform.

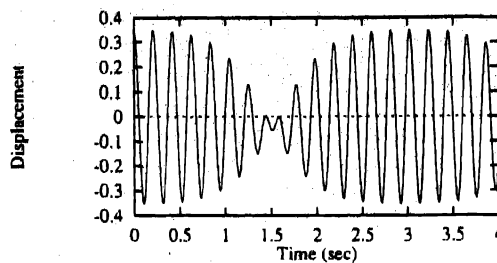


Fig. 12. Plant response, modal coupling, with desired symmetry.

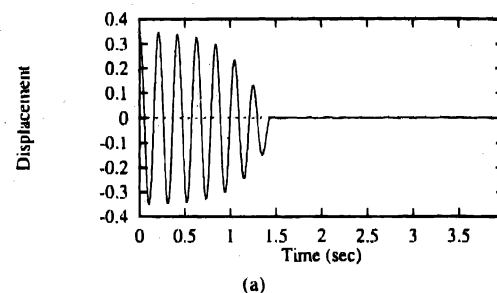


Fig. 13. DIM controller using modal coupling. (a) Plant response, (b) input.

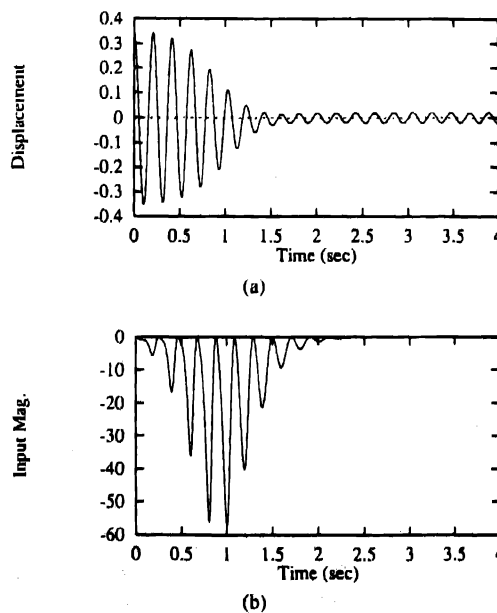


Fig. 14. DEM time invariant controller using modal coupling. (a) Plant response, (b) input.

IV. CONCLUSIONS

In this paper, new techniques were outlined to suppress the oscillations of a canonical second-order system. In general, the oscillation to be suppressed may result from the application of another control technique or may be disturbance induced.

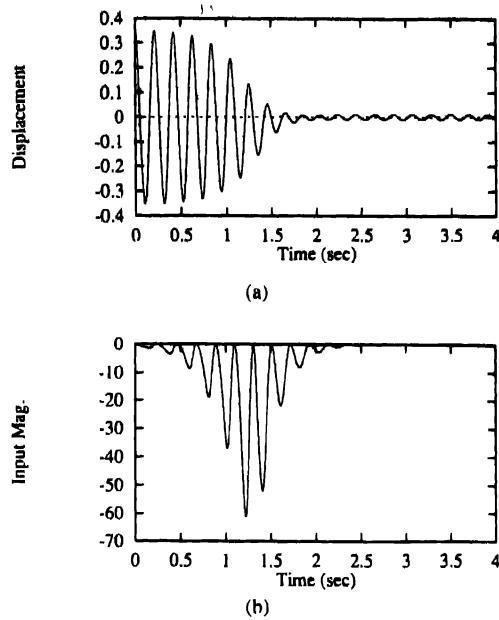


Fig. 15. DEM time-varying controller using modal coupling. (a) Plant response, (b) input.

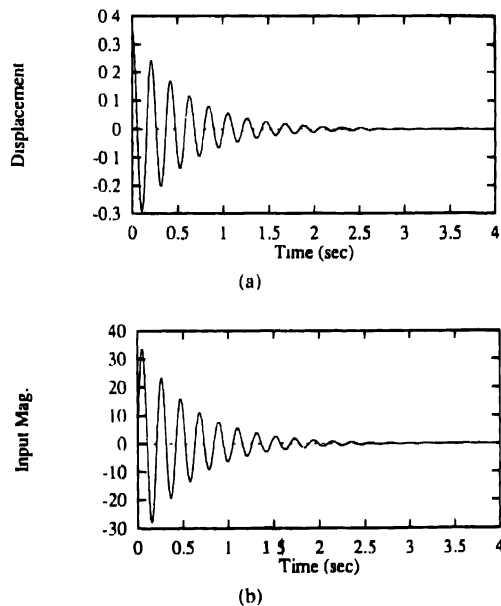


Fig. 16. PD controller. (a) Plant response, (b) input.

There are several advantages affiliated with these techniques. First of all, a great deal of intuition can accompany the mathematics in the formulation of the design as a result of the energy transfer analogy. This leads to more efficient controllers and renders troubleshooting relatively easy to perform. Secondly, the plant response approaches the operating point on the order of the cosine function whereas most existing techniques render responses that approach the operating point exponentially. Finally, the nonlinear feedback paradigm proposed in this paper yields either a positive or negative acting control input which may be useful for certain applications.

The work presented in this paper represents the first stage in the development of generalized vibration suppression laws using linear and nonlinear coupling effects. The extension to the multi-degree-of-freedom case involves the transformation of the plant equations to a form that facilitates the repeated application of the SDOF theory.

TABLE I
COORDINATE COUPLING CONTROLLER PARAMETERS

	Figure #				
	5	6	7	8	9
K	0.0017	0.005	0.005	0.0219	0.005
n	N/A	425	425	203	425
$x_1(0)$	0.35	0.35	0.35	0.35	0.35
E_1^{DIM}	N/A	N/A	0.130	N/A	N/A
ζ_c	0	0	0	0.108	0.25
E_1^{DEM}	N/A	N/A	N/A	N/A	2.09

TABLE II
MODAL COUPLING CONTROLLER PARAMETERS

	Figure #			
	10	11	12	13
K_1	0.00050	0.00050	0.00053	0.00053
γ_1, γ_2	0.5	0.75	0.75	0.75
$x_1(0)$	0.35	0.35	0.35	0.35
$x_2(0)$	0.106	0.106	0.108	0.108
E_1^{DIM}	N/A	N/A	N/A	0.0563

TABLE III

	Figure #	
	14	15
K_1	0.0005	0.0005
γ_1, γ_2	1.75	0.75
$x_1(0)$	0.35	0.35
$x_2(0)$	0.106	0.106
ζ_c	0.150	0.166
E_1^{DEM}	N/A	22.21

Thus, the concepts presented here are integral to the development of the MDOF theory. The details of the extension are contained in [15] and [16].

ACKNOWLEDGMENT

The authors wish to thank A. Duquette for his contributions and the members of the ConStruct Group, especially Dr. G. Heppler, for their suggestions and input.

REFERENCES

- [1] C. T. Chen, *Linear System Theory and Design*. New York: Holt, Rinehart and Winston, 1984.
- [2] J. M. Maciejowski, *Multivariable Feedback Design*. New York: Addison-Wesley, 1989.
- [3] M. Vidyasagar, "Optimal rejection of persistent bounded disturbances," *IEEE Trans. Automat. Contr.*, vol. AC-31, no. 6, pp. 527-534, June 1986.
- [4] J. J. E. Slotine and W. Li, *Applied Nonlinear Control*. Englewood Cliffs, NJ: Prentice-Hall, 1991.
- [5] J. C. Willems, "Dissipative dynamical systems," *Arch. Rational Mechanics Anal.*, vol. 45, pp. 321-393, 1972.
- [6] J. S. Shamma and M. Athans, "Analysis of gain scheduled control for nonlinear plants," *IEEE Trans. Automat. Contr.*, vol. 34, no. 8, pp. 898-907, Aug. 1990.
- [7] B. Kosko, *Neural Networks and Fuzzy Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1992.
- [8] M. F. Golnaraghi, "Regulation of flexible structures via nonlinear coupling," *J. Dynamics Contr.*, vol. 1, pp. 405-428, 1991.
- [9] M. F. Golnaraghi, K. L. Tuer, and D. Wang, "Regulation of flexible structures via internal resonance using nonlinear enhancement," *J. Dynamics Contr.*, vol. 4, pp. 73-96, 1994.
- [10] K. L. Tuer, A. P. Duquette, and M. F. Golnaraghi, "Vibration control of a flexible beam using a rotational internal resonance controller part I: Theoretical development and analysis," *J. Sound Vibration*, vol. 163, no. 3, 1993.
- [11] A. P. Duquette, K. L. Tuer, and M. F. Golnaraghi, "Vibration control of a flexible beam using a rotational internal resonance controller part II: Experiment," *J. Sound Vibration*, vol. 163, no. 3, 1993.
- [12] A. P. Duquette, "An experimental study of vibration control of a flexible beam via modal and coordinate coupling," Master's thesis, Univ. of Waterloo, Waterloo, Ontario, Canada, 1991.
- [13] K. L. Tuer, "Vibration control of flexible structures using nonlinear and linear coupling effects," Master's thesis, Univ. of Waterloo, Waterloo, Ontario, Canada, 1991.
- [14] K. L. Tuer, M. F. Golnaraghi, and D. Wang, "Development of a generalised active vibration suppression strategy for a cantilever beam using internal resonance," *J. Nonlinear Dynamics*, vol. 5, no. 2, pp. 131-151, 1994.
- [15] K. L. Tuer, "Towards the formulation of generalised vibration suppression laws using linear and nonlinear coupling paradigms," Ph.D. dissertation, Dept. Elec. Comp. Eng., Univ. of Waterloo, Waterloo, Ontario, Canada, 1994.
- [16] K. L. Tuer, M. F. Golnaraghi, and D. Wang, "Multi-mode vibration suppression of oscillatory systems via coupling effects," *IEEE Trans. Automat. Contr.*, submitted.
- [17] A. H. Nayfeh and D. T. Mook, *Nonlinear Oscillations*. New York: Wiley, 1979.

 μ -Synthesis of an Electromagnetic Suspension System

Masayuki Fujita, Toru Namerikawa,
Fumio Matsumura, and Kenko Uchida

Abstract—This paper deals with μ -synthesis of an electromagnetic suspension system. First, an issue of modeling a real physical electromagnetic suspension system is discussed. We derive a nominal model as well as a set of models in which the real system is assumed to reside. Different model structures and possible model parameter values are fully employed to determine unstructured additive plant perturbations, which directly yield uncertainty frequency weighting function. Second, based on the set of plant models, we setup robust performance control objectives. Third, we make use of the D - K iteration approach for the controller design. Finally, implementing the controller with a digital signal processor, experiments are carried out. With these experimental results, we show robust performance of the designed control system.

1. INTRODUCTION

Electromagnetic suspension systems can suspend objects without any contact. The increasing use of this technology in its various forms makes the research extremely active. The electromagnetic suspension technology has already applied to magnetically levitated vehicles, magnetic bearings, and so on. Recent advances on this field are shown in [1], and [5].

Feedback control is indispensable for magnetic suspension systems, since they are essentially unstable systems. To synthesis a feedback control system, a precise mathematical model for the plant is required. It is known, however, that a design model can not always express the behavior of the real physical plant. An ideal mathematical model has various uncertainties such as parameter identification errors, unmodeled dynamics, and neglected nonlinearities. The controller is required to have robustness for stability and performance against uncertainties on the model.

Recently, μ -synthesis, which is constructed with both H_∞ synthesis and μ -analysis, has been developed for the design of robust control systems [7], [8]. Beyond the singular value specifications, the μ -synthesis technique can put both robust stability and robust performance problems in a unified framework. Applications of the μ -synthesis method have been reported in [3]-[4], and [9]. This electromagnetic suspension system is a simple SISO system, but in [3] and [10], the authors also applied the H_∞/μ synthesis to a magnetic bearing, and the effectiveness of this design method was evaluated for a MIMO system. In the case of applications of H_∞/μ control to real physical systems, it is quite important to select appropriate design parameters. These parameters construct some parts of the generalized plant, e.g., uncertainty and performance weightings.

In this paper, we will evaluate μ -synthesis methodology experimentally with a real electromagnetic suspension system. We will model the additive uncertainties and decide the frequency weighting function for uncertainty accurately and reasonably. We will show that the closed-loop system with a μ controller achieves robust performance experimentally.

Manuscript received July 2, 1993; revised May 30, 1994.

M. Fujita is with the School of Information Science, Japan Advanced Institute of Science and Technology, Hokuriku Tatsunokuchi, Ishikawa 923-12, Japan.

T. Namerikawa and F. Matsumura are with the Department of Electrical and Computer Engineering, Kanazawa University 2-40-20 Kodatsuno, Kanazawa 920, Japan.

K. Uchida is with the Department of Electrical Engineering, Waseda University 3-4-1 Okubo, Shinjuku, Tokyo 169, Japan.

IEEE Log Number 9407229.

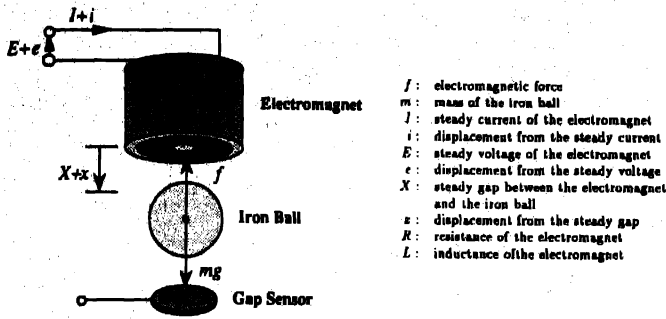


Fig. 1. Schematic diagram of the electromagnetic suspension system.

II. EXPERIMENTAL SETUP

A. Electromagnetic Suspension System

The structure of the electromagnetic suspension system is shown schematically in Fig. 1. The objective of our control experiments is to suspend an iron ball stably and firmly without any contact by controlling the attractive forces of an electromagnet. Note that this system is essentially unstable.

In Fig. 1, a cylindrical electromagnet as an actuator is located at the upper part of the experimental system. Mass of the iron ball is 1.75 kg, and it has a diameter of 77 mm. A gap sensor of our own producing is placed at the bottom of the system to measure the gap length between the iron ball and the electromagnet. The sensor is scaled for a gap of 2.4 mm per volt. It is a standard induction probe of eddy-current type. Physical parameters of this experimental machine are shown in Table I.

B. Digital Controller

The experimental machine is controlled by a digital controller using a DSP (digital signal processor). The experimental setup basically consists of the DSP which is sandwiched between A/D and D/A converters. Real-time control is implemented with a processor NEC μ PD77230, which can execute one instruction in 150 ns with 32-bit floating point arithmetic. This device has enough fast processing speed to stabilize a relatively simple magnetic suspension system in Fig. 1. The control algorithm is written in the assembly language for the DSP and a software development is assisted by a host personal computer NEC PC-9801 under the MS-DOS environment. The data acquisition board MSP-77230 consists of a 12-bit A/D converter and a 12-bit D/A converter with the maximum conversion speed of 10.5 μ s and 1.5 μ s, respectively.

The sensor outputs are filtered through an analog low-pass circuit and then converted to digital signals by A/D converters. The DSP calculates the control input signals. These digital signals are converted to analog signals by D/A converters with a range of ± 5 V. The converted signals and the steady current signals are added and amplified by 10 times to actuate the electromagnet. Steady-state voltage of the electromagnet is 24.6 V, and the maximum voltage of a regulated DC power supply is 70.0 V.

III. MODEL OF ELECTROMAGNETIC SUSPENSION SYSTEM

Our purpose in this section is to introduce an ideal mathematical model and an uncertainty weighting function for the system. See [4] for details.

TABLE I
PARAMETERS OF ELECTROMAGNETIC SUSPENSION SYSTEM

Parameter	Maximum Value	Nominal Value	Minimum Value
m [kg]	---	1.75	---
X [m]	5.50×10^{-3}	5.00×10^{-3}	4.50×10^{-3}
I [A]	1.18	1.06	0.93
x [m]	5.00×10^{-4}	0	-5.00×10^{-4}
i [A]	1.18×10^{-1}	0	-1.26×10^{-1}
L [H]	5.57×10^{-1}	5.08×10^{-1}	4.65×10^{-1}
R [Ω]	$2.37 \times 10^{+1}$	$2.32 \times 10^{+1}$	$2.27 \times 10^{+1}$
k [Nm ² /A ²]	3.35×10^{-4}	2.90×10^{-4}	2.53×10^{-4}
x_0 [m]	-3.32×10^{-4}	-6.41×10^{-4}	-9.42×10^{-4}
Q [Hm]	6.70×10^{-4}	5.79×10^{-4}	5.06×10^{-4}
X_- [m]	-3.32×10^{-4}	-6.41×10^{-4}	-9.42×10^{-4}
L_0 [H]	3.96×10^{-1}	3.75×10^{-1}	3.54×10^{-1}

A. Model Structures

We will employ four different model structures for the system depicted in Fig. 1. All of the models are finite-dimensional, linear, and time-invariant of the following state-space form

$$\begin{aligned} \dot{x} &= Ax + Bu, & y &= Cx \\ x &= [x \quad \dot{x} \quad i]^T, & u &= e, & y &= x. \end{aligned} \quad (3.1)$$

First, we introduce ideal mathematical models for the real electromagnetic suspension system. Due to the idealizing assumptions that we make, two types of ideal mathematical models can be derived hereafter, which are composed of nonlinear differential equations. We define them as Type[A] and Type[B], respectively.

Since the behavior of the electromagnetic force is nonlinear, we then employ the linearization procedure around an operating point. To account for the neglected nonlinearity, we derive two types of linear model, respectively. Thus, we will derive four linear models according to the following manners:

- *Model[A1]*: $L = \text{CONSTANT}$; and the nonlinearity of the electromagnetic forces are approximated up to the first-order term in the Taylor series expansion.
- *Model[A2]*: $L = \text{CONSTANT}$; and the nonlinearity of the electromagnetic forces are approximated up to the second-order term in the Taylor series expansion.
- *Model[B1]*: $L = L(x)$; and the nonlinearity of the electromagnetic forces are approximated up to the first-order term in the Taylor series expansion.
- *Model[B2]*: $L = L(x)$; and the nonlinearity of the electromagnetic forces are approximated up to the second-order term in the Taylor series expansion.

1) *Ideal Mathematical Model—Type[A]*: We will derive ideal mathematical models for the real electromagnetic suspension system, where the following assumptions on the electromagnet are considered.

- A.1) Magnetic permeability of the electromagnet is infinity.
- A.2) Magnetic flux density and magnetic field have not hysteresis, and they are not saturated.
- A.3) Eddy current in the magnetic pole can be neglected.

Using A.1) and A.2), we can treat the coil inductance L as a function of variable x . Then, the system can be written by the following nonlinear differential equations

$$\begin{aligned} m \frac{d^2 x}{dt^2} &= mg - f, & f &= k \left(\frac{i}{x + x_0} \right)^2 \\ e &= Ri = \frac{d}{dt} \{L(x)i\} \end{aligned} \quad (3.2)^*$$

where the coefficients k and x_0 in (3.2) are constants determined by identification experiments. Further, we introduce another assumption for Type[A].

A.A) The coil inductance is constant near an operating point. Furthermore, the electromotive forces due to the differential of gap can be neglected.

Then from (3.2), we get

$$e = Ri + L_c \frac{di}{dt}. \quad (3.3)$$

The ideal mathematical model: Type[A] is represented by (3.2) and (3.3).

Model[A1]: In view of (3.2) and (3.3), we can obtain the linear model (3.4)

$$\begin{aligned} A &= \begin{bmatrix} 0 & 1 & 0 \\ \frac{2kI^2}{m(X+x_0)^3} & 0 & -\frac{2kI}{m(X+x_0)^2} \\ 0 & 0 & -\frac{R}{L} \end{bmatrix}, \\ B &= \begin{bmatrix} 0 \\ 0 \\ \frac{1}{L} \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}^T \end{aligned} \quad (3.4)$$

Model[A2]: We can further obtain another linear model (3.5)

$$\begin{aligned} A &= \begin{bmatrix} 0 & 1 & 0 \\ \frac{2kI^2}{m(X+x_0)^3} \Delta y & 0 & -\frac{2kI}{m(X+x_0)^2} \Delta y \\ 0 & 0 & -\frac{R}{L} \end{bmatrix} \\ B &= \begin{bmatrix} 0 \\ 0 \\ \frac{1}{L} \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}^T \\ \Delta x &= \frac{x}{X+x_0}, \quad \Delta i = \frac{i}{I} \\ \Delta y &= 1 - \frac{3}{2} \Delta x + \frac{1}{2} \Delta i. \end{aligned} \quad (3.5)$$

In this way, we deal with the deviation x and i as fixed numbers, at the second-order term in the Taylor series expansion and include them in the matrix A as Δx , Δi and Δy .

2) Ideal Mathematical Model—Type[B]: For the ideal mathematical model Type[B], we also consider the assumptions A.1), A.2), A.3) and here in addition to them, we introduce the next assumption A.B) instead of A.A). Using this assumption, we can obtain more accurate model than one of Type[A].

A.B) The coil inductance L is a function of a gap x , and written as follows

$$L(x) = \frac{Q}{x + X_\infty} = L_0 \quad (3.6)$$

where the coefficients Q , X_∞ and L_0 are also the constants determined by identification experiments. For any given current i in a coil with inductance L , the magnetic co-energy is shown as $\frac{1}{2} Li^2$. Hence electromagnetic forces between the electromagnet and the iron ball in (3.2) is equal to the change rate of co-energy with respect to the distance x , i.e.,

$$\begin{aligned} f &= \frac{\partial}{\partial x} \left\{ \frac{1}{2} L(x) i^2 \right\} = \frac{1}{2} i^2 \frac{\partial L(x)}{\partial x} \\ &= \frac{Q}{2} \left(\frac{i}{x + X_\infty} \right)^2. \end{aligned} \quad (3.7)$$

Comparing (3.2) with (3.7)

$$X_\infty = x_0, \quad Q = 2k. \quad (3.8)$$

Then from (3.2), (3.6) and (3.8), we get

$$e = Ri - \frac{2ki}{(x+x_0)^2} \frac{dx}{dt} + \left(\frac{2k}{x+x_0} + L_0 \right) \frac{di}{dt}. \quad (3.9)$$

Now we obtained the ideal mathematical model: Type[B] which is constructed with (3.2) and (3.9).

Model[B1]: From (3.2) and (3.9), the linear model (3.10) (as shown at the bottom of the page) is derived.

Model[B2]: Moreover, the linear model (3.11) can be derived, as shown at the bottom of the page.

Thus, now we obtained four linear model structures: Model[A1], Model[A2], Model[B1], and Model[B2].

$$\begin{aligned} A &= \begin{bmatrix} 0 & 1 & 0 \\ \frac{2kI^2}{m(X+x_0)^3} & 0 & -\frac{2kI}{m(X+x_0)^2} \\ 0 & \frac{2kI}{(X+x_0)\{2k+L_0(X+x_0)\}} & -\frac{R(X+x_0)}{2k+L_0(X+x_0)} \end{bmatrix}, \\ B &= \begin{bmatrix} 0 \\ 0 \\ \frac{X+x_0}{2k+L_0(X+x_0)} \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}^T \end{aligned} \quad (3.10)$$

$$\begin{aligned} A &= \begin{bmatrix} 0 & 1 & 0 \\ \frac{2kI^2}{m(X+x_0)^3} \Delta y & 0 & -\frac{2kI}{m(X+x_0)^2} \Delta y \\ 0 & \frac{2kI(1-2\Delta x+\Delta i)}{(X+x_0)\{2k(1-\Delta x)+L_0(X+x_0)\}} & -\frac{R(X+x_0)}{2k(1-\Delta x)+L_0(X+x_0)} \end{bmatrix}, \\ B &= \begin{bmatrix} 0 \\ 0 \\ \frac{X+x_0}{L_0(X+x_0)+2k(1-\Delta x)} \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}^T. \end{aligned} \quad (3.11)$$

TABLE II
PERTURBED MODELS

Perturbed Model	Model Structure	Parameter Change
model(1a)	Model[A1]	$k \rightarrow k_{\max}$
model(1b)	Model[A1]	$k \rightarrow k_{\min}$
model(2a)	Model[A1]	$x_0 \rightarrow x_{0\max}$
model(2b)	Model[A1]	$x_0 \rightarrow x_{0\min}$
model(3a)	Model[A1]	$R \rightarrow R_{\max}$
model(3b)	Model[A1]	$R \rightarrow R_{\min}$
model(4a)	Model[A1]	$L \rightarrow L_{\max}$
model(4b)	Model[A1]	$L \rightarrow L_{\min}$
model(5a)	Model[A2]	$x' \rightarrow x'_{\max}$
model(5b)	Model[A2]	$x' \rightarrow x'_{\min}$
model(6a)	Model[A2]	$i' \rightarrow i'_{\max}$
model(6b)	Model[A2]	$i' \rightarrow i'_{\min}$
model(7a)	Model[B1]	$k \rightarrow k_{\max}$
model(7b)	Model[B1]	$k \rightarrow k_{\min}$
model(8a)	Model[B1]	$x_0 \rightarrow x_{0\max}$
model(8b)	Model[B1]	$x_0 \rightarrow x_{0\min}$
model(9a)	Model[B1]	$R \rightarrow R_{\max}$
model(9b)	Model[B1]	$R \rightarrow R_{\min}$
model(10a)	Model[B1]	$L_0 \rightarrow L_{0\max}$
model(10b)	Model[B1]	$L_0 \rightarrow L_{0\min}$
model(11a)	Model[B2]	$x' \rightarrow x'_{\max}$
model(11b)	Model[B2]	$x' \rightarrow x'_{\min}$
model(12a)	Model[B2]	$i' \rightarrow i'_{\max}$
model(12b)	Model[B2]	$i' \rightarrow i'_{\min}$

B. Model Parameters

To account for unpredictable perturbations in the model parameters, we will set the nominal value as well as the possible max./min. value of each parameter in every linear model. To obtain the possible max./min. value of each parameter, consider the steady-state gap $X = 5.0$ mm (nominal). Now let us perturb it with $X = 4.5$ mm and $X = 5.5$ mm (perturbed ± 0.5 mm). And, for these cases, we measured the three sets of the parameter values. The results of measurements are shown in Table II.

C. Nominal Model

We will form the nominal model using the simplest Model[A1] structure and the nominal model parameter ($X = 5.0$ mm case). Its state-space form is then of the following form

$$A_{nom} = \begin{bmatrix} 0 & 1 & 0 \\ 4481 & 0 & -18.43 \\ 0 & 0 & -45.69 \end{bmatrix}, \quad B_{nom} = \begin{bmatrix} 0 \\ 0 \\ 1.969 \end{bmatrix}, \quad C_{nom} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}^T. \quad (3.12)$$

And the corresponding nominal transfer function is

$$G_{nom} = \frac{-36.27}{(s + 66.94)(s - 66.94)(s + 45.69)}. \quad (3.13)$$

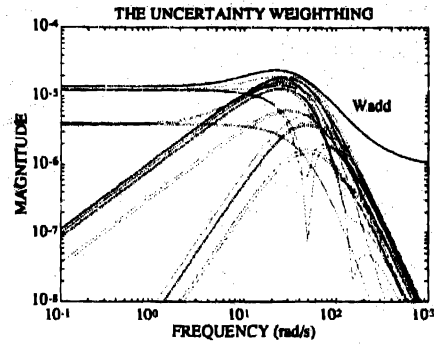


Fig. 2. Uncertainty weighting.

D. Modeling Unstructured Uncertainty

To account for unstructured uncertainties, we should consider not only a nominal model but also a set of plant models in which the real system is assumed to reside. Considering only unstructured uncertainties, we get all unstructured uncertainties together into one-full block uncertainty.

To estimate the quantities of additive model perturbations, we employ differences of gain between the nominal transfer function and the perturbed transfer function with only one parameter changed and the others fixed, where we did not consider that plural parameters change together. In such a way, 24 perturbed models have been employed. They are shown in Table II. With these notations, we can define the corresponding perturbed transfer functions \hat{G}_{ij} in an obvious way

$$\Delta_{ij} := \hat{G}_{ij} - G_{nom} \quad (1 \leq i \leq 12, j = a, b). \quad (3.14)$$

Frequency responses of these additive perturbations $|\Delta_{ij}(j\omega)|$ are plotted in Fig. 2, with 24 dotted lines. Now let us consider the set of plant models. Here we assume the following form

$$G := \{G_{nom} + \Delta_{add}W_{add} : \|\Delta_{add}\|_{\infty} \leq 1\} \quad (3.15)$$

in which the real plant is assumed to reside. All of the uncertainties are captured in the normalized, unknown transfer function Δ_{add} . It is natural to choose the uncertainty weighting W_{add} as follows (shown in Fig. 2). Here it should be noted that the magnitude of the uncertainty weighting W_{add} covers all the model perturbations shown in Fig. 2

$$W_{add} = \frac{1.4 \times 10^{-5} (1 + s/8)(1 + s/170)(1 + s/420)}{(1 + s/30)(1 + s/35)(1 + s/38)}. \quad (3.16)$$

IV. DESIGN

A. Control Objectives

Electromagnetic suspension system is essentially unstable. We must design a robust controller to stabilize the closed-loop system; furthermore, we would like to design a controller to maintain the performance against unpredictable disturbances and the uncertainties.

Let us consider the feedback structure shown in Fig. 3. The box represents the set of the models: G of the real system. Robust stability requirement for the additive uncertainty can be evaluated using the closed-loop transfer function $K S$, where $S := (I + GK)^{-1}$. Hence robust stability test for $G \in \mathcal{G}$ is equivalent to

$$\|W_{add} K (I + G_{nom} K)^{-1} W_{add}\|_{\infty} < 1. \quad (4.1)$$

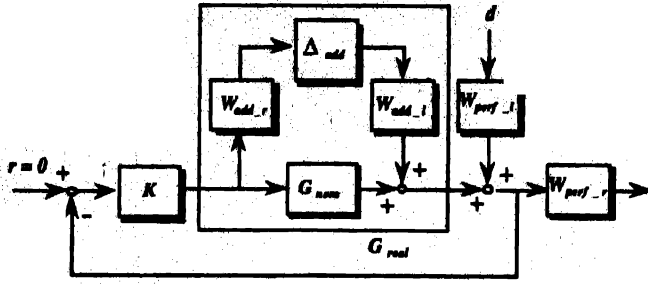


Fig. 3. Feedback structure.

It is noted in Fig. 3 that we factor the uncertainty weighting as $W_{add} = W_{add_l} \times W_{add_r}$, where

$$W_{add_l} = 1.0 \times 10^{-5},$$

$$W_{add_r} = \frac{1.4 \times (1 + s/8)(1 + s/170)(1 + s/420)}{(1 + s/30)(1 + s/35)(1 + s/38)}. \quad (4.2)$$

To reject the disturbances at low frequency band, the performance weighting W_{perf} is now chosen as

$$W_{perf} = \frac{200.0}{1 + s/0.1}. \quad (4.3)$$

We also factor the performance weighting as $W_{perf} = W_{perf_l} \times W_{perf_r}$, where

$$W_{perf_l} = 1.0 \times 10^{-5}, \quad W_{perf_r} = \frac{2.0 \times 10^7}{1 + s/0.1}. \quad (4.4)$$

In practical situation, however, we would like to achieve this performance specification for all the possible plant $G \in \mathbf{G}$. A necessary and sufficient condition for this robust performance is

$$\|W_{perf_r}(I + GK)^{-1}W_{perf_l}\|_{\infty} < 1, \quad \forall G \in \mathbf{G}. \quad (4.5)$$

Now the control objective is to find a stabilizing controller K which achieves the following two conditions

- The closed-loop system remains internally stable for every plant model $G \in \mathbf{G}$,
- The weighted sensitivity function satisfies the performance test (4.5) for every plant $G \in \mathbf{G}$.

The design objectives have been specified as the requirements for particular closed loop transfer functions with the frequency weighting functions W_{add} and W_{perf} . The above control objectives exactly fit in the μ -synthesis framework by introducing a fictitious uncertainty block Δ_{perf} . Rearranging the feedback structure in Fig. 3, we can build the interconnection structure shown in Fig. 4.

B. μ -Synthesis

We first define a block structure Δ_P as

$$\Delta_P := \left\{ \begin{bmatrix} \Delta_{add} & 0 \\ 0 & \Delta_{perf} \end{bmatrix} : \Delta_{add} \in \mathbf{C}, \Delta_{perf} \in \mathbf{C} \right\}. \quad (4.6)$$

Next, consider a generalized plant P partitioned as

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}. \quad (4.7)$$

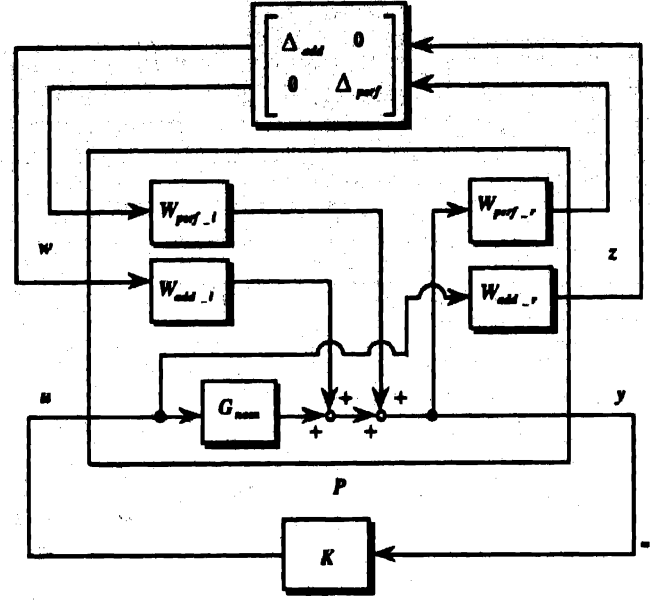


Fig. 4. Interconnection structure.

Obviously in Fig. 4, we can get a lower linear fractional transformation $\mathcal{F}_l(P, K)$ on P by K

$$\mathcal{F}_l(P, K) := P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21}. \quad (4.8)$$

Finally, robust performance condition is equivalent to the following structured singular value μ test

$$\sup_{\omega \in \mathbf{R}} \mu_{\Delta_P}(\mathcal{F}_l(P, K)(j\omega)) < 1. \quad (4.9)$$

The complex structured singular value μ_{Δ_P} is defined as

$$\mu_{\Delta_P}(M) := \frac{1}{\min \{ \bar{\sigma}(\Delta) : \Delta \in \Delta, \det(I - M\Delta) = 0 \}} \quad (4.10)$$

unless no $\Delta \in \Delta$ makes $I - M\Delta$ singular, in which case $\mu_{\Delta}(M) := 0$. In this case a matrix M in (4.10) belongs to $\mathbf{C}^{2 \times 2}$.

C. D-K iteration

Unfortunately, it is not known how to obtain a controller K achieving the structured singular value test (4.9) directly. But we can obtain the lower and upper bounds of μ . Our approach taken here is the so-called D - K iteration procedure.

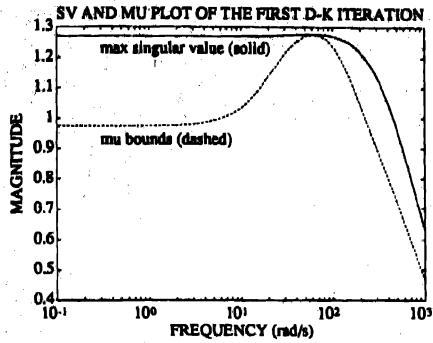
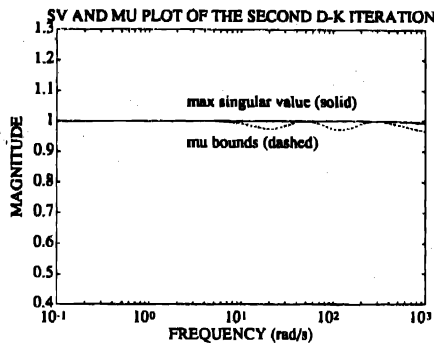
The D - K iteration involves a sequence of minimizations over either K or D while holding the other fixed, until a satisfactory controller is constructed. First, for $D = I$ fixed, the controller K_1 is synthesized using the well-known state-space H_{∞} optimization method. Let $P_1 = P$ denote the given open-loop interconnection structure in Fig. 4, and $\mathcal{F}_l(P, K)$ be the closed-loop transfer function from the disturbances w to the errors z .

Then, solving the following H_{∞} control problem

$$\|\mathcal{F}_l(P_1, K_1)\|_{\infty} < \gamma_1, \quad \gamma_1 = 1.3. \quad (4.11)$$

The problem (4.11) yields the central controller K_1 as shown in (4.12) found at the bottom of the page.

$$K_1 = \frac{-5.22 \times 10^8 (s + 12.46)(s + 30.0)(s + 35.0)(s + 38.0)(s + 45.69)(s + 66.94)}{(s + 0.10)(s + 31.6 - j5.12)(s + 31.6 + j5.12)(s + 39.77)(s + 315.2 - j329.6)(s + 315.2 + j329.6)(s + 734.7)} \quad (4.12)$$

Fig. 5. σ and μ plot of the first D - K iteration.Fig. 6. σ and μ plot of the second D - K iteration.

Here we try to assess robust performance of this closed-loop system using μ -analysis associated with the block structure (4.6). The maximum singular value and μ upper bound of the closed-loop transfer function $\mathcal{F}_l(P_1, K_1)$ are plotted in Fig. 5. It is noteworthy to point out that the peak value of the upper bound μ plot is not less than one. This reveals that the closed-loop system with this H_∞ controller K_1 does not achieve robust performance condition.

Next, the above calculations of μ produce a scaling matrix at each frequency. In this design, we try to fit the curve using a first-order transfer function.

Now, let P_2 denote the new open-loop interconnection structure absorbing the scaling matrix D . This time, from the following H_∞ control problem

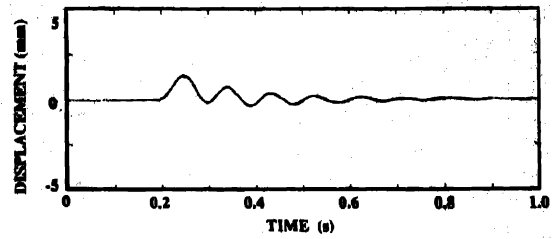
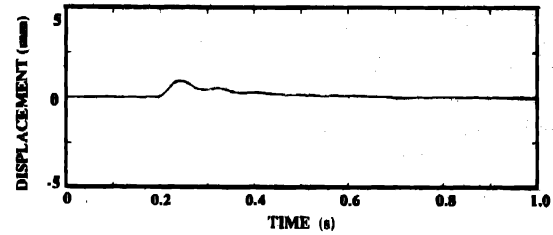
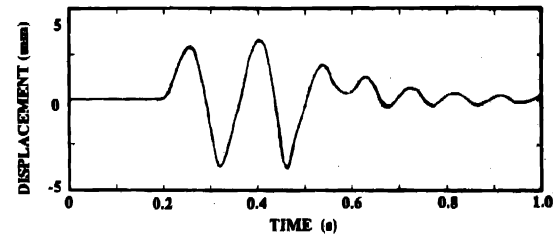
$$\|\mathcal{F}_l(P_2, K_2)\|_\infty < \gamma_2, \quad \gamma_2 = 1.0 \quad (4.13)$$

we can calculate the controller K_2 as found in (4.14) as shown at the bottom of the page.

The maximum singular value and μ upper bound of this closed-loop system are plotted in Fig. 6. Since the value of μ is less than one in Fig. 6, robust performance condition is now achieved.

V. EXPERIMENTAL RESULTS

The designed controllers K_1 and K_2 are continuous-time systems. To implement these two controllers with the digital controller, we discretized them via the well known Tustin transform. The controllers K_1 and K_2 are discretized at the sampling period of $45\mu\text{s}$ and $60\mu\text{s}$, respectively.

Fig. 7. Response to step disturbance with K_1 (-17.15N).Fig. 8. Response to step disturbance with K_2 (-17.15N).Fig. 9. Response to step disturbance with K_1 (-34.30N).

We succeeded in the stable suspension of the iron ball using both of the controllers K_1 and K_2 . In the Section IV, robust stability and robust performance objectives were considered as the control problems. The obtained H_∞ controller K_1 achieves robust stability condition, and μ controller K_2 achieves not only robust stability but also robust performance specification. Hence, we will evaluate robust performance as well as robust stability of the closed-loop systems with responses against various external disturbances.

There the disturbances are added to the experimental system as an applied voltage in the electromagnet. It is noted that there are four types of disturbances. Taking account that the steady-state force of the electromagnet is equal to 17.15 N, we added the following disturbance forces to the floating iron ball

$$\text{downward } 17.15 \text{ N}, \quad \text{downward } 34.30 \text{ N}.$$

These disturbances are large enough to evaluate the robustness of both these two controllers. Experimental results are shown in Figs. 7–10.

First of all, these experimental results in Figs. 7–10 show that the iron ball is suspended. Responses in Fig. 9 are vibrating extremely, however, their vibration get on the decrease. This shows the closed-loop systems with both the controllers K_1 and K_2 remain stable against these disturbances. Comparing Fig. 7 with Fig. 9, the responses with K_1 deteriorate extremely against relatively large disturbances. While in Fig. 8 and Fig. 10, the responses with the controller K_2 maintain good transient responses against these disturbances. Now we can see the following observation.

$$K_2 = \frac{-8.01 \times 10^9 (s + 10.54)(s + 15.75)(s + 30.0)(s + 35.0)(s + 38.0)}{(s + 0.10)(s + 19.59 - j5.32)(s + 19.59 + j5.32)(s + 38.48 - j2.70)(s + 38.48 + j2.70)} \\ \times \frac{(s + 45.69)(s + 66.94)(s + 169.6)}{(s + 176.6)(s + 420.1 - j272.8)(s + 420.1 + j272.8)(s + 8180)} \quad (4.14)$$

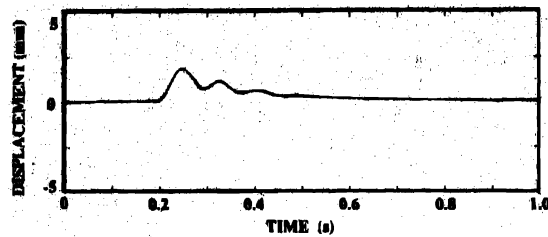


Fig. 10. Response to step disturbance with K_2 ($-34.30N$).

- The closed-loop system with the μ controller K_2 achieves robust performance, while the closed-loop system with the H_∞ controller K_1 does not.

VI. CONCLUSIONS

In this paper, we experimentally evaluated a controller designed by μ -synthesis methodology with an electromagnetic suspension system. We have obtained a nominal mathematical model as well as a set of plant models in which the real system is assumed to reside. With this set of the models we designed the control system to achieve robust performance objective utilizing μ -synthesis method.

First, four types of different model structures were derived based on the several idealizing assumptions for the real system. Second, for every model, the nominal value as well as the possible maximum and minimum values of each model parameter was determined by measurements and/or experiments. Third, a nominal model was naturally chosen. This model has the simplest model structure of all four models and makes use of nominal parameter values. Then, model perturbations were defined to account for additive unstructured uncertainties from such as neglected nonlinearities and model parameter errors. Fourth, we defined a family of plant models where the unstructured additive perturbation was employed. The method to model the plant as belonging to a family or set plays a key role for systematic robust control design. Fifth, we setup robust performance objective as a structured singular value test. Next, for the design, the D - K iteration approach was employed. Finally, the experimental results showed that the closed-loop system with the μ -controller achieves not only nominal performance and robust stability, but in addition robust performance.

REFERENCES

- [1] P. Allaire, Ed., "Magnetic bearings," in *Proc. Third Int. Symp. Magnetic Bearings*, Alexandria, VA, 1992.
- [2] G. J. Balas, P. Young, and J. C. Doyle, "The process of control design for the NASA Langley minimag structure," in *Proc. Amer. Contr. Conf.*, Boston, MA, 1991, pp. 562-567.
- [3] M. Fujita, K. Hatake, F. Matsumura, and K. Uchida "An experimental evaluation and comparison of H_∞/μ control for a magnetic bearing," in *Proc. 12th IFAC World Congress*, Sydney, Australia, 1993, pp. 393-398.
- [4] M. Fujita, T. Namerikawa, F. Matsumura, and K. Uchida, " μ -synthesis of an electromagnetic suspension system," in *Proc. 31st IEEE Conf. Decis. Contr.*, Tucson, AZ, 1992, pp. 2574-2579.
- [5] T. Higuchi, "Magnetic bearings," in *Proc. Second Int. Symp. Magnetic Bearings*, Tokyo, Japan, 1990.
- [6] F. Matsumura and S. Tachimori, "Magnetic suspension system suitable for wide range operation (in Japanese)," *Trans. IEE of Japan*, vol. 99-B, pp. 25-32, 1978.
- [7] A. Packard and J. Doyle, "The complex structured singular value," *Automatica*, vol. 29, no. 1, pp. 71-109, 1993.
- [8] G. Stein and J. C. Doyle, "Beyond singular values and loop shapes," *J. Guidance*, vol. 14, no. 1, pp. 5-16 1991.

- [9] M. Steinbuch, G. Schotstra, and O. H. Bosgra, "Robust control of a compact disc player," in *Proc. IEEE Conf. Decis. Contr.*, Tucson, AZ, 1992, pp. 2596-2600.
- [10] M. Fujita, K. Hatake, and F. Matsumura, "Loop shaping based robust control of a magnetic bearing," *IEEE Contr. Syst. Mag.*, vol. 13, no. 4, pp. 57-65, Aug. 1993.

Parameter-Dependent Lyapunov Functions and the Popov Criterion in Robust Analysis and Synthesis

Wassim M. Haddad and Dennis S. Bernstein

Abstract—Many practical applications of robust feedback control involve constant real parameter uncertainty, whereas small gain or norm-bounding techniques guarantee robust stability against complex, frequency-dependent uncertainty, thus entailing undue conservatism. Since conventional Lyapunov bounding techniques guarantee stability with respect to time-varying perturbations, they possess a similar drawback. In this paper we develop a framework for parameter-dependent Lyapunov functions, a less conservative refinement of "fixed" Lyapunov functions. An immediate application of this framework is a reinterpretation of the classical Popov criterion as a parameter-dependent Lyapunov function. This result is then used for robust controller synthesis with full-order and reduced-order controllers.

I. INTRODUCTION

The analysis and synthesis of robust feedback controllers entails a fundamental distinction between parametric and nonparametric uncertainty. Parametric uncertainty refers to plant uncertainty that is modeled as constant real parameters, whereas nonparametric uncertainty refers to uncertain transfer function gains modeled as complex frequency-dependent quantities. In the time domain, nonparametric uncertainty is manifested as time-varying uncertain real parameters.

The distinction between parametric and nonparametric uncertainty is critical to the achievable performance of feedback control systems. For example, in the problem of vibration suppression for flexible space structures, if stiffness matrix uncertainty is modeled as nonparametric uncertainty, then perturbations to the damping matrix will inadvertently be allowed. Predictions of stability and performance for given feedback gains will consequently be extremely conservative, thus limiting achievable performance [1]. Alternatively, this problem can be viewed by considering the classical analysis of Hill's equation (e.g., the Mathieu equation) which shows that time-varying parameter variations can destabilize a system even when the parameter variations are confined to a region in which constant variations are nondestabilizing. Consequently, a feedback controller designed for time-varying parameter variations will unnecessarily sacrifice performance when the uncertain real parameters are actually constant.

Manuscript received August 9, 1991; revised July 28, 1992, August 20 1993, and May 30, 1994. This work was supported in part by Air Force Office of Scientific Research Grant F49620-92-J-0127 and National Science Foundation Grant ECS-9109558.

W. M. Haddad is with School of Aerospace Engineering Georgia Institute of Technology Atlanta, GA 30332-0150 USA.

D. S. Bernstein is with Department of Aerospace Engineering The University of Michigan Ann Arbor, MI 48109-2118 USA.

IEEE Log Number 9407728.

The above distinction can also be illustrated by considering the central result of feedback control theory, namely, the small gain theorem, which guarantees robust stability by requiring that the loop gain (including desired weighting functions for loop shaping) be less than unity at all frequencies. The small gain theorem, however, does not make use of phase information in guaranteeing stability. In fact, the small gain theorem allows the loop transfer function to possess arbitrary phase at all frequencies, although in many applications at least some knowledge of phase is available [2]. Thus, small gain techniques such as H_∞ theory are generally conservative when phase information is available. More generally, since $|e^{j\omega}| = 1$ regardless of the phase angle ϕ , it can be expected that any robustness theory based upon norm bounds will suffer from the same shortcoming. Of course, every real parameter can be viewed as a complex parameter with phase angle $\phi = 0$ degrees or $\phi = 180$ degrees.

To some extent, phase information is accounted for by means of positivity theory [3]–[15] which is widely used to model passive systems such as flexible structures [16], [17]. In this theory, a positive real plant and strictly positive real uncertainty are both assumed to have phase less than 90 degrees so that the loop transfer function has less than 180 degrees of phase shift, hence guaranteeing robust stability in spite of gain uncertainty. Both gain and phase properties can be simultaneously accounted for by means of the circle criterion [15], [18]–[22] which yields the small gain theorem and positivity theorem as special cases. It is important to note that since positivity theory and the circle criterion can be obtained from small gain conditions by means of suitable transformations, they can be viewed as equivalent results from a mathematical point of view. The engineering ramifications of the ability to include phase information, however, can be significant [1].

The above discussion is further illuminated by means of Lyapunov function theory in [15]. Specifically, as pointed out in [15], a serious defect of conventional or fixed Lyapunov bounding theory is the fact that stability is guaranteed even if the plant uncertainty ΔA is a function of t . This observation follows from the fact that the Lyapunov derivative $\dot{V}(x(t)) = V_x(x(t))(A + \Delta A(t))x(t)$ need only be negative for each fixed value of t [15], [23]. Although this feature is desirable if ΔA is time varying, as discussed above, it leads to conservatism when ΔA is actually constant. This defect can be remedied, however, by utilizing an alternative approach, which is consistent with Lyapunov function bounding techniques, based upon parameter-dependent Lyapunov functions. The idea behind parameter-dependent Lyapunov functions is to allow the Lyapunov function to be a function of the uncertainty ΔA . In the usual case, $V(x) = x^T P x$, P is a single, fixed matrix, whereas the parameter-dependent Lyapunov function $V_{\Delta A}(x) = x^T P(\Delta A)x$ represents a family of Lyapunov functions.

The concept of a parameter-dependent Lyapunov functions is not new to this paper. Specifically, a parameter-dependent Lyapunov function of the form $V(x) = x^T P(\lambda_1, \dots, \lambda_r)x$, where $P(\lambda_1, \dots, \lambda_r) = \sum_{i=1}^r \lambda_i P_i$, is considered in [24]. In this case the matrices P_i correspond to the vertices of a polytope of uncertain matrices with vertices A_1, \dots, A_r . More recently, [25] considers a Lyapunov function with matrix $P(\sigma_1, \dots, \sigma_s) = P_0 + \sum_{i=1}^s \sigma_i P_i$, where P_0 corresponds to the nominal system and the P_i are "first-order perturbations" of P_0 . Numerical techniques are used to determine P_i and the range of robust stability. Both [24] and [25] discuss potential advantages of parameter-dependent Lyapunov functions over fixed Lyapunov functions.

The goal of the present paper is to develop robust analysis and synthesis techniques that exploit the fact that the classical Popov criterion [26] is based upon a parameter-dependent Lyapunov

function. Indeed, recall that the Popov criterion is based upon the Lur'e-Postnikov Lyapunov function

$$V_\phi(x) = x^T P x + N \int_0^y \phi(\sigma) d\sigma \quad (1.22)$$

where $y = Cx$ and $\phi(\cdot)$ is a scalar memoryless time-invariant nonlinearity in the sector $[0, k]$, that is, $0 \leq \phi(y)y \leq ky^2$. Specializing to the linear uncertainty case $\phi(y) = Fy$, where $0 \leq F \leq k$, yields

$$\begin{aligned} V_F(x) &= x^T P x + N \int_0^y F \sigma d\sigma = x^T P x + NF \frac{y^2}{2} \\ &= x^T [P + \frac{1}{2} N F C^T C] x = x^T P(F)x. \end{aligned}$$

This form appears in [10, pp. 84–89] and was discussed in the context of robust analysis in [15].

For practical purposes the form of the parameter-dependent Lyapunov function $V_F(x)$ is useful since the presence of F restricts the allowable time-varying uncertain parameters [27]. That is, if $F(t)$ were permitted, then terms involving $\dot{F}(t)$ would arise and potentially subvert the negative definiteness of $\dot{V}_F(x)$.

This paper has four specific goals: 1) to provide a general framework for parameter-dependent Lyapunov functions; 2) to obtain a generalized multivariable version of the Popov criterion for linear matrix uncertainty ΔA (the classical Popov criterion is limited to scalar or diagonal nonlinearities) along with H_2 robust performance bounds; 3) to provide explicit uncertainty bounds for the multivariable Popov criterion in terms of a single Riccati equation that can be used for robust controller synthesis; and 4) to develop robust controller synthesis techniques based upon the multivariable Popov criterion with applications to full-order and reduced-order controllers.

Notation.

- $R, R^{* \times}, R'$ —real numbers, $r \times s$ real matrices, $R^{* \times 1}$
- $C, C^{* \times}, C'$ —complex numbers, $r \times s$ complex matrices, $C^{* \times 1}$
- $E, \text{tr}, 0, \times, \dots$ —expectation, trace, $r \times s$ zero matrix, Kronecker product
- $I, (\cdot)^T, (\cdot)^*$ — $r \times r$ identity, transpose, complex conjugate transpose
- $(\cdot)^{-T}, (\cdot)^{-*}$ —inverse transpose, complex conjugate inverse transpose
- S', N', P' — $r \times r$ symmetric, nonnegative-definite, positive-definite matrices
- $Z_1 \leq Z_2, Z_1 < Z_2$ — $Z_2 - Z_1 \in N', Z_2 - Z_1 \in P', Z_1, Z_2 \in S'$
- $\|Z\|_1, \|G(s)\|_2$ — $[\text{tr} Z Z^*]^{1/2}, [(1/2\pi) \int_{-\infty}^{\infty} \|G(j\omega)\|_F^2 d\omega]^{1/2}$

II. ROBUST STABILITY AND PERFORMANCE PROBLEMS: ANALYSIS

Let $\mathcal{U} \subset \mathbb{R}^{n \times n}$ denote a set of perturbations ΔA of a given nominal dynamics matrix $A \in \mathbb{R}^{n \times n}$. Within the context of robustness analysis, it is assumed that A is asymptotically stable and $0 \in \mathcal{U}$. We begin by considering the stability of $A + \Delta A$ for all $\Delta A \in \mathcal{U}$.

Robust Stability Problem: Determine whether the linear system

$$\dot{x}(t) = (A + \Delta A)x(t), \quad t \in [0, \infty) \quad (2.1)$$

is asymptotically stable for all $\Delta A \in \mathcal{U}$.

To consider the problem of robust performance, we introduce an external disturbance model involving white noise signals as in standard LQG (H_2) theory. The robust performance problem concerns the worst-case H_2 norm, that is, the worst-case over \mathcal{U} of the expected value of a quadratic form involving outputs $z(t) = Ex(t)$, where $E \in \mathbb{R}^{q \times n}$, when the system is subjected to a standard white noise disturbance $w(t) \in \mathbb{R}^d$ with weighting $D \in \mathbb{R}^{n \times d}$.

Robust Performance Problem: For the disturbed linear system

$$\dot{x}(t) = (A + \Delta A)x(t) + Dw(t), \quad t \in [0, \infty). \quad (2.2)$$

$$z(t) = Ex(t) \quad (2.3)$$

where $w(\cdot)$ is a zero-mean d -dimensional white noise signal with intensity I_d , determine a performance bound β satisfying

$$J(\mathcal{U}) \triangleq \sup_{\Delta A \in \mathcal{U}} \limsup_{t \rightarrow \infty} \mathbb{E}\{\|z(t)\|_2^2\} \leq \beta. \quad (2.4)$$

In Section VI, (2.2) will denote a control system in closed-loop configuration subjected to external white noise disturbances and for which $z(t)$ denotes the state and control regulation error.

Next, we express the H_2 performance measure in terms of the observability Gramian for the pair $(A + \Delta A, E)$. For convenience define the $n \times n$ nonnegative-definite matrices $R \triangleq E^T E$, $V \triangleq DD^T$.

Lemma 2.1: Suppose $A + \Delta A$ is asymptotically stable for all $\Delta A \in \mathcal{U}$. Then

$$J(\mathcal{U}) = \sup_{\Delta A \in \mathcal{U}} \text{tr} P_{\Delta A} V = \sup_{\Delta A \in \mathcal{U}} \|G_{\Delta A}(s)\|_2^2 \quad (2.5)$$

where $P_{\Delta A} \in \mathbb{R}^{n \times n}$ is the unique, nonnegative-definite solution to

$$0 = (A + \Delta A)^T P_{\Delta A} + P_{\Delta A} (A + \Delta A) + R \quad (2.6)$$

and $G_{\Delta A}(s) \triangleq E[sI - (A + \Delta A)]^{-1}D$.

III. ROBUST STABILITY AND PERFORMANCE VIA PARAMETER-DEPENDENT LYAPUNOV FUNCTIONS

The key step in obtaining robust stability and performance is to bound the uncertain terms $\Delta A^T P_{\Delta A} + P_{\Delta A} \Delta A$ in the Lyapunov equation (2.6) by means of a parameter-dependent bounding function $\Omega(P, \Delta A)$ which guarantees robust stability by means of a family of Lyapunov functions. This procedure corresponds to the construction of a parameter-dependent Lyapunov function which constrains the class of allowable time-varying uncertainties. The following result forms the basis for all later developments.

Theorem 3.1. Let $\Omega_0: \mathbb{N}^n \rightarrow \mathbb{S}^n$ and $P_0: \mathcal{U} \rightarrow \mathbb{S}^n$ be such that

$$\Delta A^T P + P \Delta A \leq \Omega_0(P) - [(A + \Delta A)^T P_0(\Delta A) + P_0(\Delta A)(A + \Delta A)], \quad \Delta A \in \mathcal{U}, P \in \mathbb{N}^n \quad (3.1)$$

and suppose there exists $P \in \mathbb{N}^n$ satisfying

$$0 = A^T P + P A + \Omega_0(P) + R \quad (3.2)$$

and such that $P + P_0(\Delta A)$ is nonnegative definite for all $\Delta A \in \mathcal{U}$. Then

$$(A + \Delta A, E) \text{ is detectable, } \Delta A \in \mathcal{U} \quad (3.3)$$

if and only if

$$A + \Delta A \text{ is asymptotically stable, } \Delta A \in \mathcal{U}. \quad (3.4)$$

In this case

$$P_{\Delta A} \leq P + P_0(\Delta A), \quad \Delta A \in \mathcal{U} \quad (3.5)$$

where $P_{\Delta A}$ is given by (2.6). Therefore

$$J(\mathcal{U}) \leq \text{tr} P V + \sup_{\Delta A \in \mathcal{U}} \text{tr} P_0(\Delta A) V. \quad (3.6)$$

If, in addition, there exists $\tilde{P}_0 \in \mathbb{S}^n$ such that

$$P_0(\Delta A) \leq \tilde{P}_0, \quad \Delta A \in \mathcal{U} \quad (3.7)$$

then

$$J(\mathcal{U}) \leq \text{tr}[(P + \tilde{P}_0)V]. \quad (3.8)$$

Proof: Note that in (3.1), P denotes an arbitrary element of \mathbb{N}^n , whereas in (3.2) P denotes a specific solution of the modified Lyapunov equation (3.2). This minor abuse of notation considerably simplifies the presentation. Now, note that for all $\Delta A \in \mathbb{R}^{n \times n}$, (3.2) is equivalent to

$$0 = (A + \Delta A)^T P + P(A + \Delta A) + \Omega_0(P) - (\Delta A^T P + P \Delta A) + R. \quad (3.9)$$

Adding and subtracting $(A + \Delta A)^T P_0(\Delta A) + P_0(\Delta A)(A + \Delta A)$ to (3.9) yields

$$\begin{aligned} 0 = & (A + \Delta A)^T (P + P_0(\Delta A)) + (P + P_0(\Delta A))(A + \Delta A) \\ & + \Omega_0(P) - [(A + \Delta A)^T P_0(\Delta A) + P_0(\Delta A)(A + \Delta A)] \\ & - (\Delta A^T P + P \Delta A) + R. \end{aligned} \quad (3.10)$$

Hence, by assumption, (3.10) has a solution $P \in \mathbb{N}^n$ for all $\Delta A \in \mathbb{R}^{n \times n}$. If ΔA is restricted to the set \mathcal{U} , then, by (3.1), $\Omega_0(P) - [(A + \Delta A)^T P_0(\Delta A) + P_0(\Delta A)(A + \Delta A)] - (\Delta A^T P + P \Delta A)$ is nonnegative definite. Thus if condition (3.3) holds for all $\Delta A \in \mathcal{U}$, then Theorem 3.6 of [28] implies $(A + \Delta A, [R + \Omega(P, \Delta A) - (\Delta A^T P + P \Delta A)]^{1/2})$ is detectable for all $\Delta A \in \mathcal{U}$, where

$$\Omega(P, \Delta A) \triangleq \Omega_0(P) - [(A + \Delta A)^T P_0(\Delta A) + P_0(\Delta A)(A + \Delta A)]. \quad (3.11)$$

It now follows from (3.10) and Lemma 12.2 of [28] that $A + \Delta A$ is asymptotically stable for all $\Delta A \in \mathcal{U}$. Conversely, if $A + \Delta A$ is asymptotically stable for all $\Delta A \in \mathcal{U}$, then (3.3) is immediate. Now, subtracting (2.6) from (3.10) yields

$$\begin{aligned} 0 = & (A + \Delta A)^T (P + P_0(\Delta A) - P_{\Delta A}) + (P + P_0(\Delta A) - P_{\Delta A})(A + \Delta A) \\ & \times (A + \Delta A) + \Omega_0(P) - [(A + \Delta A)^T P_0(\Delta A) + P_0(\Delta A)(A + \Delta A)] \\ & \times (A + \Delta A) - (\Delta A^T P + P \Delta A), \quad \Delta A \in \mathcal{U} \end{aligned} \quad (3.12)$$

or, equivalently, since $A + \Delta A$ is asymptotically stable for all $\Delta A \in \mathcal{U}$

$$\begin{aligned} P + P_0(\Delta A) - P_{\Delta A} &= \int_0^\infty e^{(A + \Delta A)^T t} [\Omega(P, \Delta A) \\ &\quad - (\Delta A^T P + P \Delta A)] e^{(A + \Delta A) t} dt \\ &\geq 0, \quad \Delta A \in \mathcal{U} \end{aligned} \quad (3.13)$$

which implies (3.5). The performance bounds (3.6) and (3.8) are now an immediate consequence of (2.5), (3.5), and (3.7) \square

Note that with $\Omega(P, \Delta A)$ defined by (3.11) condition (3.1) can be written as

$$\Delta A^T P + P \Delta A \leq \Omega(P, \Delta A), \quad \Delta A \in \mathcal{U}, \quad P \in \mathbb{N}^n \quad (3.1)'$$

where $\Omega(P, \Delta A)$ is a function of the uncertainty ΔA . For convenience we shall say that $\Omega(\cdot, \cdot)$ is a parameter-dependent Ω -bound, which is consistent with [29]. One can recover the standard guaranteed cost bound or parameter-independent Ω -bound by setting $P_0(\Delta A) \equiv 0$ so that $\Omega(P, \Delta A) \equiv \Omega_0(P)$ and therefore $\Delta A^T P + P \Delta A \leq \Omega_0(P)$ for all $\Delta A \in \mathcal{U}$. Finally, since we do not assume that $P_0(0) = 0$, it follows that $\Omega_0(P)$ need not be nonnegative definite. If, however, $P_0(0) = 0$, then it follows from (3.1) with $\Delta A = 0$ that $\Omega_0(P) \geq 0$ for all nonnegative-definite P . To apply Theorem 3.1, we first specify functions $\Omega_0(\cdot)$ and $P_0(\cdot)$ and an uncertainty set \mathcal{U} such that (3.1)' holds. If the existence of a nonnegative-definite solution P to (3.2) can be determined analytically or numerically and the detectability condition (3.3) is satisfied, then robust stability is guaranteed and the performance bound (3.8) can be computed.

Finally, we show that a parameter-dependent Ω -bound establishing robust stability is equivalent to the existence of a parameter-dependent Lyapunov function which also establishes robust stability. To show this, assume there exists a positive-definite solution to (3.2), let

$P_0: \mathcal{U} \rightarrow \mathbb{N}^n$, and define the parameter-dependent Lyapunov function $V_{\Delta A}(x) \triangleq x^T (P + P_0(\Delta A))x$. Note that since P is positive definite and $P_0(\Delta A)$ is nonnegative definite, $V_{\Delta A}(x)$ is positive definite. The corresponding Lyapunov derivative is given by

$$\dot{V}_{\Delta A}(x) = -x^T [\Omega_0(P) - \{\Delta A^T P + P \Delta A + (A + \Delta A)^T \times P_0(\Delta A) + P_0(\Delta A)(A + \Delta A)\} + R]x. \quad (3.14)$$

Thus, using (3.1) it follows that $\dot{V}_{\Delta A}(x) \leq 0$ so that $A + \Delta A$ is stable in the sense of Lyapunov. Asymptotic stability follows from the invariant set theorem.

IV. CONSTRUCTION OF PARAMETER-DEPENDENT LYAPUNOV FUNCTIONS

We now assign explicit structure to the uncertainty set \mathcal{U} and the parameter-dependent bounding function $\Omega(\cdot)$. Specifically, let

$$\mathcal{U} \triangleq \{\Delta A \in \mathbb{R}^{n \times n} \mid \Delta A = B_0 F C_0, F \in \mathcal{F}\} \quad (4.1)$$

where \mathcal{F} satisfies

$$\mathcal{F} \subseteq \mathcal{F} = \{F \in \mathbb{R}^{m_0 \times m_0} \mid 0 \leq F \leq M\} \quad (4.2)$$

and where $B_0 \in \mathbb{R}^{n \times m_0}$, $C_0 \in \mathbb{R}^{m_0 \times n}$ are fixed matrices denoting the structure of the uncertainty, $F \in \mathbb{R}^{m_0 \times m_0}$ is an uncertain symmetric matrix, and $M \in \mathbb{R}^{m_0 \times m_0}$ is a given positive-definite matrix. Note that \mathcal{F} may be equal to \mathcal{F} , although, for generality, \mathcal{F} may be a specified proper subset of \mathcal{F} . For example, \mathcal{F} may consist of block-structured matrices $F = \text{block-diag}(F_1, F_2, \dots, F_r)$ where $F_i = \text{block-diag}(F_{i1}, F_{i2}, \dots, F_{im_i})$ and $M = \text{block-diag}(M_1, M_2, \dots, M_r)$, then $0 \leq F_i \leq M_i$, $i = 1, \dots, r$. Finally, we assume that $0 \in \mathcal{F}$ and $M \in \mathcal{F}$.

Next, we provide an equivalent characterization of the set \mathcal{F} .

Lemma 4.1 Let $F \in \mathbb{S}^{m_0}$ and $M \in \mathbb{P}^{m_0}$. Then $F M^{-1} F \leq I$ if and only if $0 \leq F \leq M$.

For \mathcal{U} given by (4.1), the parameter-dependent bound $\Omega(\cdot)$ satisfying (3.12) can now be given a concrete form. Since the elements ΔA in \mathcal{U} are parameterized by the elements F in \mathcal{F} , we shall write $P_0(F)$ in place of $P_0(\Delta A)$. Finally, we define the sets \mathcal{N}_s and \mathcal{N}_{nd} such that the product of the transpose of every matrix in \mathcal{N}_s (resp., \mathcal{N}_{nd}) and every matrix in \mathcal{F} is symmetric (respectively, nonnegative definite) by

$$\mathcal{N}_s \triangleq \{N \in \mathbb{R}^{m_0 \times m_0} \mid F N = N^T F, F \in \mathcal{F}\}$$

and

$$\mathcal{N}_{nd} \triangleq \{N \in \mathcal{N}_s \mid F N \geq 0, F \in \mathcal{F}\}$$

Finally, Lemma 4.1 of [30] implies that there exists $\mu \in \mathbb{N}^{m_0}$ such that $F N \leq \mu$ for all $F \in \mathcal{F}$.

Proposition 4.1 Let $N \in \mathcal{N}_s$ and

$$(M^{-1} - N C_0 B_0) + (M^{-1} - N C_0 B_0)^T > 0 \quad (4.3)$$

Furthermore, let \mathcal{U} be defined by (4.1) and define $\Omega_0(\cdot)$ and $P_0(\cdot)$ by

$$\Omega_0(P) = (C_0 + N C_0 A + B_0^T P)^T [(M^{-1} - N C_0 B_0) + (M^{-1} - N C_0 B_0)^T]^{-1} (C_0 + N C_0 A + B_0^T P) \quad (4.4)$$

and

$$P_0(F) = C_0^T F N C_0 \quad (4.5)$$

Then (3.1) is satisfied

Proof Since by (4.3) $(M^{-1} - N C_0 B_0) + (M^{-1} - N C_0 B_0)^T$ is positive definite and by Lemma 4.1 $F - F M^{-1} F$ is nonnegative definite, it follows that

$$\begin{aligned} 0 &\leq [(C_0 + N C_0 A + B_0^T P) - [(M^{-1} - N C_0 B_0) \\ &\quad + (M^{-1} - N C_0 B_0)^T] F C_0]^T [(M^{-1} - N C_0 B_0) \\ &\quad + (M^{-1} - N C_0 B_0)^T]^{-1} [(C_0 + N C_0 A + B_0^T P) \\ &\quad - [(M^{-1} - N C_0 B_0) + (M^{-1} - N C_0 B_0)^T] F C_0] \\ &\quad + 2 C_0^T [F - F M^{-1} F] C_0 \\ &= \Omega_0(P) - C_0^T F (C_0 + N C_0 A + B_0^T P) \\ &\quad - (C_0^T + A^T C_0^T N^T + P B_0) F C_0 + C_0^T F [(M^{-1} - N C_0 B_0) \\ &\quad + (M^{-1} - N C_0 B_0)^T] F C_0 + 2 C_0^T [F - F M^{-1} F] C_0 \\ &= \Omega_0(P) - C_0^T F B_0^T P - P B_0 F C_0 \\ &\quad - C_0^T F N C_0 A - A^T C_0^T N^T F C_0 \\ &\quad - C_0^T F N C_0 B_0 F C_0 - C_0^T F B_0^T C_0^T N^T F C_0 \\ &= \Omega_0(P) - [(A + \Delta A)^T P_0(F) + P_0(F)(A + \Delta A)] \\ &\quad - [\Delta A^T P + P \Delta A] \end{aligned}$$

which proves (3.1) with \mathcal{U} given by (4.1) \square

Next, using Theorem 3.1 and Proposition 4.1 we have the following immediate result

Theorem 4.1 Let $N \in \mathcal{N}_{nd}$ and assume (4.3) is satisfied. Furthermore, suppose there exists a nonnegative-definite matrix P satisfying

$$0 = A^T P + P A + (C_0 + N C_0 A + B_0^T P)^T [(M^{-1} - N C_0 B_0) + (M^{-1} - N C_0 B_0)^T]^{-1} (C_0 + N C_0 A + B_0^T P) + R \quad (4.6)$$

Then

$$(A + \Delta A, E) \text{ is detectable, } \Delta A \in \mathcal{U} \quad (4.7)$$

if and only if

$$A + \Delta A \text{ is asymptotically stable, } \Delta A \in \mathcal{U} \quad (4.8)$$

In this case, if $\mu \in \mathbb{N}^{m_0}$ satisfies $F N \leq \mu$ for all $F \in \mathcal{F}$, then

$$J(\mathcal{U}) \leq \text{tr}[(P + C_0^T \mu C_0) V] \quad (4.9)$$

Proof The result is a direct specialization of Theorem 3.1 using Proposition 4.1 with $P_0(\Delta A) = C_0^T F N C_0$. Since by assumption, $F N \geq 0$ for all $F \in \mathcal{F}$, it follows that $P + P_0(F)$ is nonnegative definite for all $F \in \mathcal{F}$ as required by Theorem 3.1 \square

Note that asymptotic stability in Theorem 4.1 is guaranteed by the parameter-dependent Lyapunov function $V_{\Delta A}(x) = x^T (P + C_0^T F N C_0)x$.

Remark 4.1 The condition $F N = N^T F, F \in \mathcal{F}$ is analogous to the commuting assumption between the D -scales and Δ blocks in μ -analysis which accounts for structure in the uncertainty F . Note that there always exists such a matrix N even if $F \in \mathcal{F}$ is not diagonal. For example, if $F = F_1 I_{m_0}$, where F_1 is a scalar uncertainty, then N can be an arbitrary symmetric matrix. Alternatively, if F is nondiagonal, then one can choose $N = N_0 I_{m_0}$, where N_0 is a scalar. Finally, F and N may be block diagonal with commuting blocks situated on the diagonal. Characterization of the optimal multiplier N for robust controller analysis and synthesis is given in Section VI.

Remark 4.2 Standard loop-shifting techniques [31] can be used to consider uncertainties with upper and lower bounds of the form $M_1 \leq F \leq M_2$, where $F \in \mathcal{F}$ and $M_1, M_2 \in \mathbb{S}^{m_0}$. In this case, Proposition 4.1 holds with F, A , and M replaced by $F - M_1, A + B_0 M_1 C_0$, and $M_2 - M_1$, respectively. Similar modifications can be made to Theorem 4.1.

Next, we use results from positivity theory to guarantee the existence of a positive-definite solution to (4.6). Let $G(s) \sim \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ denote a state space realization of a transfer function $G(s)$, that is, $G(s) = C(sI - A)^{-1}B + D$. The notation " \sim " denotes a minimal realization:

Lemma 4.2. (Positive Real Lemma [4], [9]):

$$G(s) \sim \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

is positive real if and only if there exist matrices P, L , and W with P positive definite such that

$$0 = A^T P + PA + L^T L, \quad (4.10)$$

$$0 = PB - C^T + L^T W, \quad (4.11)$$

$$0 = D + D^T - W^T W. \quad (4.12)$$

Next, we show that if $D + D^T > 0$ then (4.10)-(4.12) yield a single Riccati equation characterizing positive realness. For the statement of this result recall that a square transfer function $G(s)$ is strongly positive real if it is strictly positive real [15] and $D + D^T > 0$.

Lemma 4.3 [14]. Let $G(s) \sim \begin{bmatrix} A & B \\ C & D \end{bmatrix}$. Then $G(s)$ is strongly positive real if and only if $D + D^T > 0$ and there exist positive-definite matrices P and R such that

$$0 = A^T P + PA + (C - B^T P)^T (D + D^T)^{-1} (C - B^T P) + R. \quad (4.13)$$

We now use Lemma 4.3 to obtain a sufficient condition for the existence of a solution to (4.6).

Theorem 4.2. Let

$$\mathcal{G}(s) \sim \begin{bmatrix} A & -B_0 \\ C_0 + NC_0 A & M^{-1} - NC_0 B_0 \end{bmatrix}.$$

Then $\mathcal{G}(s)$ is strongly positive real if and only if there exist positive definite matrices P and R satisfying (4.6).

Finally, we specialize Proposition 4.1 and Theorem 4.1 to the case in which $N = 0$ and $M = D_0^{-1}$, where $D_0 + D_0^T > 0$. In this case we have the following result.

Proposition 4.2: Let $D_0 \in \mathbb{R}^{m_0 \times m_0}$ be such that $D_0 + D_0^T > 0$. Furthermore, let \mathcal{U} be defined by (4.1) with $M = D_0^{-1}$, let $P_0(F) = 0$, and define $\Omega_0(\cdot)$ by

$$\Omega_0(P) = (C_0 + B_0^T P)^T (D_0 + D_0^T)^{-1} (C_0 + B_0^T P). \quad (4.14)$$

Then (3.1) is satisfied.

Since $P_0(F) = 0$, the case $N = 0$ corresponds to a parameter-independent Ω -bound. Hence, it follows from Theorem 3.1 that if there exists a nonnegative-definite matrix P satisfying

$$0 = A^T P + PA + (C_0 + B_0^T P)^T (D_0 + D_0^T)^{-1} (C_0 + B_0^T P) + R \quad (4.15)$$

then $(A + \Delta A, E)$ is detectable for all $\Delta A \in \mathcal{U}$ if and only if $A + \Delta A$ is asymptotically stable for all $\Delta A \in \mathcal{U}$. Furthermore, it follows from Lemma 4.3 that the existence of a positive-definite matrix P satisfying (4.15) implies that

$$G(s) \sim \begin{bmatrix} A & -B_0 \\ C_0 & D_0 \end{bmatrix}$$

is strongly positive real. Hence the parameter-independent Ω -bound (4.14) guarantees robust stability in the presence of positive real (but otherwise unknown) plant uncertainty. The situation is analogous to H_∞ bounded real theory, which also depends upon a parameter independent Ω -bound.

V. CONNECTIONS TO THE POPOV CRITERION

In this section we demonstrate connections between the parameter-dependent Lyapunov function obtained in Section 4 and the classical multivariable Popov criterion. Traditionally, the Popov criterion is stated for component-decoupled time-invariant sector-bounded nonlinearities $\phi(y)$. We state the Popov criterion for this case and then specialize to the case of linear uncertainty. Hence let $M \in \mathbb{R}^{m_0 \times m_0}$ be a given positive-definite matrix and define

$$\Phi \triangleq \{ \phi : \mathbb{R}^{m_0} \rightarrow \mathbb{R}^{m_0} : \phi^T(y)[M^{-1}\phi(y) - y] \leq 0, y \in \mathbb{R}^{m_0}, \text{ and } \phi(y) = [\phi_1(y_1), \phi_2(y_2), \dots, \phi_{m_0}(y_{m_0})]^T \}. \quad (5.1)$$

If $M = \text{diag}(M_1, \dots, M_{m_0})$ is diagonal, then the sector condition characterizing Φ is implied by the scalar sector conditions $0 \leq \phi_i(y_i)y_i \leq M_i y_i^2, y_i \in \mathbb{R}, i = 1, \dots, m_0$.

Theorem 5.1. (The Popov Criterion) [15]: Suppose there exists a nonnegative-definite matrix $N = \text{diag}(N_1, \dots, N_{m_0})$ such that $M^{-1} + (I + N\kappa)G(s)$ is strongly positive real, where $G(s) \sim \begin{bmatrix} A & B \\ C & 0 \end{bmatrix}$. Then, for all $\phi(\cdot) \in \Phi$, the negative feedback interconnection of $G(s)$ and $\phi(\cdot)$ is asymptotically stable with Lyapunov function

$$V_\phi(x) = x^T P x + 2 \sum_{i=1}^{m_0} \int_0^{y_i} \phi_i(\sigma) N_i d\sigma. \quad (5.2)$$

Next, we specialize Theorem 5.1 to the case of constant linear parameter uncertainty. Specifically, consider the system $\dot{x}(t) = (A + \Delta A)x(t)$, where $\Delta A \in \mathcal{U}$ and \mathcal{U} is defined by

$$\mathcal{U} \triangleq \{ \Delta A : \Delta A = -BFC, \quad F = \text{diag}(F_1, F_2, \dots, F_{m_0}), \quad 0 \leq F_i \leq M_i, \quad i = 1, \dots, m_0 \}.$$

By setting $\phi(y) = Fy = FCx$ Theorem 5.1 guarantees that $A + \Delta A$ is asymptotically stable for all $\Delta A \in \mathcal{U}$.

It has thus been shown that in the special case that F and N are diagonal nonnegative-definite matrices, Theorem 4.1 (with B_0 replaced by $-B_0$) specializes to the Popov criterion when applied to linear parameter uncertainty. This is not surprising since the Lyapunov function (5.2) that establishes robust stability has the form

$$V_F(x) = x^T P x + 2 \sum_{i=1}^{m_0} \int_0^{y_i} F_i \sigma N_i d\sigma, \quad y_i = (C_0 x)_i, \quad (5.3)$$

or, equivalently

$$V_F(x) = x^T P x + x^T C_0^T F N C_0 x = x^T P x + \sum_{i=1}^{m_0} F_i N_i x^T C_0^T C_0 x \quad (5.4)$$

which is a specialization of the parameter-dependent Lyapunov function considered in Section IV to the case of diagonal uncertainty F . The results of Section IV, however, allowed nondiagonal uncertain matrices F , which cannot be addressed by means of the nonlinear theory. Finally, note that the uncertain parameters F are not allowed to be arbitrarily time-varying, which is consistent with the fact that the Popov criterion is restricted to time-invariant nonlinearities.

VI. ROBUST CONTROLLER SYNTHESIS VIA PARAMETER-DEPENDENT LYAPUNOV FUNCTIONS: FIXED-ORDER DYNAMIC COMPENSATION

In this section we consider robust stability and performance with dynamic output-feedback controllers. For generality, the compensator dimension n_c may be less than the plant order n . Define $\hat{n} \triangleq n + n_c$, where $n_c \leq n$.

Dynamic Robust Stability and Performance Problem Given the n th-order stabilizable and detectable plant with constant structured real-valued plant parameter variations

$$\dot{x}(t) = (A + \Delta A)x(t) + Bu(t) + D_1w(t), t \geq 0, \quad (6.1)$$

$$y(t) = Cx(t) + D_2u(t) \quad (6.2)$$

where $u(t) \in \mathbb{R}^m$, $x(t) \in \mathbb{R}^d$, and $y(t) \in \mathbb{R}^l$, determine an n th-order dynamic compensator

$$\dot{v}(t) = A_c v(t) + B_c y(t), \quad (6.3)$$

$$u(t) = C_c v(t) \quad (6.4)$$

that satisfies the following design criteria

- i) The closed-loop system (6.1)-(6.4) is asymptotically stable for all $\Delta A \in \mathcal{U}$ and
- ii) The performance functional

$$J(A, B, C) \triangleq \sup_{\Delta A \in \mathcal{U}} \limsup_{t \rightarrow \infty} \frac{1}{t} E \left\{ \int_0^t [v^T(s) R_1 v(s) + u^T(s) R_2 u(s)] ds \right\} \quad (6.5)$$

is minimized

For each uncertain variation $\Delta A \in \mathcal{U}$, the closed-loop system (6.1)-(6.4) can be written as

$$\dot{x}(t) = (A + \Delta A)x(t) + Du(t) \quad t \geq 0 \quad (6.6)$$

where

$$x(t) = \begin{bmatrix} x(t) \\ v(t) \end{bmatrix}, A \triangleq \begin{bmatrix} A & BC \\ B_c C & A_c \end{bmatrix}, \Delta A = \begin{bmatrix} \Delta A & 0_{n \times n} \\ 0_{n \times n} & 0_{n \times n} \end{bmatrix}$$

and where the closed loop disturbance $Du(t)$ has intensity $V = DD^T$ where $D = \begin{bmatrix} D_1 \\ B, D_2 \end{bmatrix}$, $V \triangleq \begin{bmatrix} V_1 & 0 \\ 0 & B, V_2 B_c^T \end{bmatrix}$, $V_1 = D_1 D_1^T$, $V_{1,2} = D_1 D_2^T = 0$, $V_2 = D_2 D_2^T$. The closed-loop system uncertainty ΔA has the form $\Delta A = B_0 \Gamma C_0$ where $B_0 = \begin{bmatrix} B_0 \\ 0_{n_c \times m_0} \end{bmatrix}$, $C_0 \triangleq [C_0 \ 0_{m_0 \times l}]$. Finally, if $A + \Delta A$ is asymptotically stable for all $\Delta A \in \mathcal{U}$ for a given compensator (A_c, B_c, C_c) , then it follows from Lemma 2.1 that the performance functional (6.5) is given by

$$J(A, B, C) = \sup_{\Delta A \in \mathcal{U}} \text{tr} P_{\Delta A} V \quad (6.7)$$

where $P_{\Delta A}$ satisfies the $n \times n$ Lyapunov equation

$$0 = (A + \Delta A)^T P_{\Delta A} + P_{\Delta A} (A + \Delta A) + R \quad (6.8)$$

where

$$R = [E_1 \ E_2 C_c^T], \quad R = E^T F = \begin{bmatrix} R_1 & 0 \\ 0 & C_c^T R_2 C_c \end{bmatrix}$$

Next, we apply Theorem 4.1 to controller synthesis. Specifically, we replace the Lyapunov equation (6.8) for the dynamic problem with a Riccati equation that guarantees that the closed-loop system is robustly stable. Thus, for the dynamic output feedback problem, Theorem 4.1 holds with A, R, V replaced by $\hat{A}, \hat{R}, \hat{V}$. This leads to the following problem

Dynamic Auxiliary Minimization Problem Determine $N \in \mathcal{N}_{nd}$ and controllable and observable (A_c, B_c, C_c) that minimize

$$\mathcal{J}(A_c, B_c, C_c, N) \triangleq \text{tr}(P + C_0^T \mu C_0) V \quad (6.9)$$

where $P \in \mathbb{N}^n$ satisfies

$$0 = \hat{A}^T P + P \hat{A} + (C_0 + \lambda C_0 A + B_0^T P)^T [(M^{-1} - N \tilde{C}_0 B_0) + (M^{-1} - \lambda C_0 B_0)^T]^{-1} (C_0 + \lambda C_0 A + B_0^T P) + R \quad (6.10)$$

Necessary conditions for the dynamic auxiliary minimization problem will provide fixed-order dynamic output feedback controllers with guaranteed robust stability and performance. The following result is required for the statement of the main theorem

Lemma 6.1 [32] Let Q, P be $n \times n$ nonnegative-definite matrices and suppose that $\text{rank } QP = n$. Then there exist $n \times n$ matrices G, Γ and an $n_c \times n_c$ invertible matrix M , unique except for a change of basis in \mathbb{R}^{n_c} , such that

$$QP = G^T \Gamma M \quad \Gamma G^T = I_{n_c} \quad (6.11)$$

Furthermore, the $n \times n$ matrices $\tau \triangleq G^T \Gamma$ and $\tau_{\perp} \triangleq I_n - \tau$ are idempotent and have rank n_c and $n - n_c$, respectively

To state the main result of this section let $P, Q \in \mathbb{R}^{n \times n}$ and define the notation

$$R_0 \triangleq (M^{-1} - \lambda C_0 B_0) + (M^{-1} - \lambda C_0 B_0)^T$$

$$C \triangleq C_0 + \lambda C_0 A, \quad A_c \triangleq A + B_0 R_0^{-1} C,$$

$$B_{c,c} \triangleq B_2 + B^T C_0^T \lambda^T R_0^{-1} \lambda C_0 B,$$

$$P_c \triangleq B^T P + B^T C_0^T \lambda^T R_0^{-1} (C + B_0^T P)$$

$$\Sigma \triangleq C^T \lambda_2^{-1} C, \quad \lambda_P \triangleq A_P - Q \Sigma + B_0 R_0^{-1} B_0^T P$$

$$A_Q \triangleq A_c + B_0 R_0^{-1} B_0^T P - (I + B_0 R_0^{-1} \lambda C_0) B R_{2a}^{-1} P_a$$

Theorem 6.1 Let $n_c \leq n$ assume $R_0 > 0$, and assume $V \in \mathcal{N}_{nd}$. Furthermore, suppose there exist $n \times n$ nonnegative-definite matrices P, Q, P_c, Q_c satisfying

$$0 = A_P^T P + P A_P + R_1 + C^T R_0^{-1} C + P B_0 R_0^{-1} B_0^T P - P^T R_{2,c}^{-1} P_c + \tau_{\perp}^T P_c^T R_{2,c}^{-1} P_c \tau_{\perp} \quad (6.12)$$

$$0 = (A_c + B_0 R_0^{-1} B_0^T [P + P_c]) Q + Q (A_P + B_0 R_0^{-1} B_0^T [P + P_c])^T + V_1 - Q \Sigma Q + \tau_{\perp} Q \Sigma Q \tau_{\perp}^T \quad (6.13)$$

$$0 = A_c^T P + P A_c + P B_0 R_0^{-1} B_0^T P + P_c^T R_{2,c}^{-1} P_c - \tau_{\perp}^T P_c^T R_{2,c}^{-1} P_c \tau_{\perp} \quad (6.14)$$

$$0 = A_Q Q + Q A_Q^T + Q \Sigma Q - \tau_{\perp} Q \Sigma Q \tau_{\perp}^T \quad (6.15)$$

$$\text{rank } Q = \text{rank } P = \text{rank } QP = n_c \quad (6.16)$$

and let A_c, B_c, C_c be given by

$$A_c = \Gamma [A_Q - Q \Sigma] G^T, \quad B_c = \Gamma Q C^T \lambda_2^{-1}, \quad C_c = -R_{2,c}^{-1} P_c G^T \quad (6.17)$$

Then $(A + \Delta A, E)$ is detectable for all $\Delta A \in \mathcal{U}$ if and only if $A + \Delta A$ is asymptotically stable for all $\Delta A \in \mathcal{U}$. In this case the performance of the closed-loop system (6.7) satisfies the bound

$$J(A_c, B_c, C_c) \leq \text{tr}[(P + \hat{P}) V_1 + P Q \Sigma Q + C_0^T \mu C_0 V_1] \quad (6.18)$$

Proof: The proof is constructive in nature and is similar to the proofs given in [14] and [33]. Specifically, first we obtain necessary conditions for the Dynamic Auxiliary Minimization Problem and then show by construction that these conditions serve as sufficient conditions for robust stabilization and provide a worst-case H_2 performance bound. \square

Theorem 6.1 provides constructive sufficient conditions that yield dynamic feedback gains A_c, B_c, C_c for robust stability and performance. When solving (6.12)–(6.15) numerically, the matrices M and N and the structure matrices B_0 and C_0 appearing in the design equations can be adjusted to examine tradeoffs between performance and robustness. Finally, to further reduce conservatism, one can view the multiplier matrix N as a free parameter and optimize the worst-case H_2 performance bound \mathcal{J} with respect to N . In particular, setting $\partial \mathcal{J} / \partial N = 0$ yields

$$0 = \frac{1}{2} M \tilde{C}_0 \tilde{C}_0^T + [(M^{-1} - N \tilde{C}_0 \tilde{B}_0) + (M^{-1} - N \tilde{C}_0 \tilde{B}_0)^T]^{-1} \times (\tilde{C}_0 + N \tilde{C}_0 \tilde{A} + \tilde{B}_0^T \tilde{P}) \tilde{Q} \tilde{A}^T \tilde{C}_0^T + [(M^{-1} - N \tilde{C}_0 \tilde{B}_0) + (M^{-1} - N \tilde{C}_0 \tilde{B}_0)^T]^{-1} (\tilde{C}_0 + N \tilde{C}_0 \tilde{A} + \tilde{B}_0^T \tilde{P}) \tilde{Q} (\tilde{C}_0 + N \tilde{C}_0 \tilde{A} + \tilde{B}_0^T \tilde{P})^T [(M^{-1} - N \tilde{C}_0 \tilde{B}_0) + (M^{-1} - N \tilde{C}_0 \tilde{B}_0)^T]^{-1} \tilde{B}_0^T \tilde{C}_0^T \quad (6.19)$$

where \tilde{Q} satisfies

$$0 = (\tilde{A} + \tilde{B}_0 R_0^{-1} N \tilde{C}_0 \tilde{A} + \tilde{B}_0 R_0^{-1} \tilde{C}_0 + \tilde{B}_0 R_0^{-1} \tilde{B}_0^T \tilde{P}) \tilde{Q} + \tilde{Q} (\tilde{A} + \tilde{B}_0 R_0^{-1} N \tilde{C}_0 \tilde{A} + \tilde{B}_0 R_0^{-1} \tilde{C}_0 + \tilde{B}_0 R_0^{-1} \tilde{B}_0^T \tilde{P})^T + \tilde{V}. \quad (6.20)$$

By using (6.19) within a numerical search algorithm, the optimal compensator and multiplier N can be determined simultaneously, thus avoiding the need to iterate between controller design and optimal multiplier evaluation.

Remark 6.1: Several special cases can immediately be discerned from Theorem 6.1. For example, in the full-order case, set $n_c = n$ so that $\tau = G = \Gamma = I_n$ and $\tau_1 = 0$. In this case the last term in each of (6.12)–(6.15) is zero and (6.15) is superfluous. Alternatively, letting $B_0 = 0, C_0 = 0$ and retaining the reduced-order constraint $n_c \leq n$ yields the result of [32].

VII. NUMERICAL ALGORITHM AND ILLUSTRATIVE RESULTS

In this section we describe a numerical algorithm for solving the Riccati equation (6.10) along with the expression (6.19) for the optimal multiplier N . We also present numerical results for controller synthesis via an illustrative example.

To synthesize dynamic compensators, we let $\mu = (M_2 - M_1)N$ in (6.9) and determine (A_c, B_c, C_c, N) to minimize $\mathcal{J}(A_c, B_c, C_c, N)$ subject to (6.10) with $\tilde{P} \in \mathbb{N}^n$. To do this we form the Lagrangian

$$\mathcal{L}(A_c, B_c, C_c, N, \tilde{P}, \tilde{Q}) = \text{tr}[(P + \tilde{C}_0^T (M_2 - M_1) N \tilde{C}_0) \tilde{V} + \{(\tilde{A} + \tilde{B}_0 M_1 \tilde{C}_0)^T \tilde{P} + \tilde{P} (\tilde{A} + \tilde{B}_0 M_1 \tilde{C}_0) + [\tilde{C}_0 + N \tilde{C}_0 (\tilde{A} + \tilde{B}_0 M_1 \tilde{C}_0) + \tilde{B}_0^T \tilde{P}]^T \cdot \tilde{R}_0^{-1} [\tilde{C}_0 + N \tilde{C}_0 (\tilde{A} + \tilde{B}_0 M_1 \tilde{C}_0) + \tilde{B}_0^T \tilde{P}] + \tilde{R} \} \tilde{Q}] \quad (7.1)$$

where

$$\tilde{R}_0 \triangleq [(M_2 - M_1)^{-1} - N \tilde{C}_0 \tilde{B}_0] + [(M_2 - M_1)^{-1} - N \tilde{C}_0 \tilde{B}_0]^T \quad (7.2)$$

and $\tilde{Q} \in \mathbb{R}^{n \times n}$ is a Lagrange multiplier. The partial derivatives of \mathcal{L} are then used in the search procedure. Note that the shifted version of (6.10) discussed in Remark 4.2 is used in (7.1) to address uncertainties with upper and lower bounds of the form $M_1 \leq F \leq M_2$.

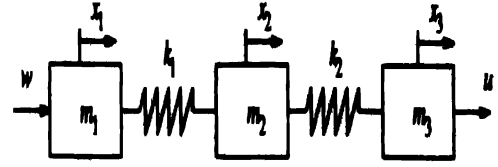


Fig. 1. Three-mass system.

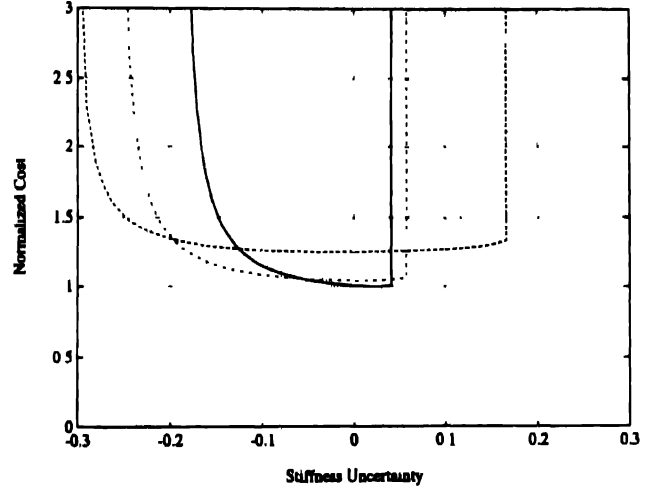


Fig. 2. Performance vs. robustness trade-off for LQG and Popov controllers

A quasi-Newton search algorithm was initialized with $N = 0$ and the LQG gains. For given values of the robustness bounds M_1 and M_2 , the search algorithm was used to find A_c, B_c, C_c and N satisfying the necessary conditions. After each iteration, M_1 and M_2 were increased and the current design was used as the initial step for the next iteration. Since the optimal compensator and multiplier are found simultaneously, there is no need to iterate between controller design and optimal multiplier evaluation.

Consider the three-mass, two-spring system shown in Fig. 1 with $m_1 = m_2 = m_3 = 1$ and an uncertain spring stiffness k_2 . A control force acts on mass 3 while the position of mass 1 is measured resulting in a noncollocated control problem. The nominal dynamics, with state variables defined in Fig. 1, are given by

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -k_1 & k_1 & 0 & 0 & 0 & 0 \\ k_1 & -(k_1 + k_{2nom}) & k_{2nom} & 0 & 0 & 0 \\ 0 & k_{2nom} & -k_{2nom} & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad C = [1 \ 0 \ 0 \ 0 \ 0 \ 0], \quad D_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$D_2 = [0 \ 1]$, and $k_1 = k_{2nom} = 1$. The actual spring stiffness of the second spring can be written as $k_2 = k_{2nom} + \Delta k$ so that the actual dynamics are given by $A(\Delta k) = A + B_0 \Delta k C_0$, where $B_0 = [0 \ 0 \ 0 \ 0 \ 1 \ 1]^T$ and $C_0 = [0 \ 1 \ 0 \ 0 \ 0 \ 0]$. Furthermore, let

$$E_1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad E_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

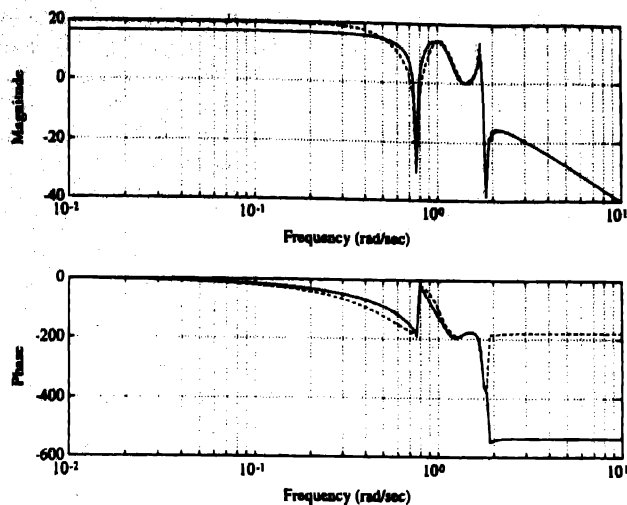


Fig. 3. Popov (dashed) and LQG (solid) controllers.

Two full-order ($n_c = n$) Popov compensators were designed. Fig. 2 compares performance versus robustness trade-offs of the Popov compensators (dashed) with the normalized LQG design (solid). Fig. 3 shows the magnitude and phase of both a Popov design and the LQG design. Note that the Popov design robustified the LQG controller notch by increasing both the width and the depth of the notch.

REFERENCES

- [1] D. S. Bernstein, W. M. Haddad, and D. C. Hyland, "Small gain versus positive real modeling of real parameter uncertainty," *AIAA J. Guid. Contr. Dyn.*, vol. 15, pp. 538-540, 1992.
- [2] D. S. Bernstein, E. G. Collins, Jr., and D. C. Hyland, "Real parameter uncertainty and phase information in the robust control of flexible structures," in *Proc. IEEE Conf. Dec. Contr.*, Honolulu, HI, Dec. 1990, pp. 379-380.
- [3] G. Zames, "On the input-output stability of time-varying nonlinear feedback systems, part I: Conditions derived using concepts of loop gain, conicity, and positivity," *IEEE Trans. Automat. Contr.*, vol. AC-11, pp. 228-238, 1966.
- [4] B. D. O. Anderson, "A system theory criterion for positive real matrices," *SIAM J. Contr. Optim.*, vol. 5, pp. 171-182, 1967.
- [5] J. C. Willems, "The generation of Lyapunov functions for input-output stable systems," *SIAM J. Contr. Optim.*, vol. 9, pp. 105-634, 1971.
- [6] B. D. O. Anderson, "The small-gain theorem, the passivity theorem, and their equivalence," *J. Franklin Inst.*, vol. 293, pp. 105-115, 1972.
- [7] J. C. Willems, "Dissipative dynamical systems part I: General theory; Part II: Quadratic supply rates," *Arch. Rat. Mech.*, vol. 45, pp. 321-351, 352-393, 1972.
- [8] V. M. Popov, *Hyperstability of Control Systems*. New York: Springer-Verlag, 1973.
- [9] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis: A Modern Systems Theory Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1973.
- [10] K. S. Narendra and J. H. Taylor, *Frequency Domain Criteria for Absolute Stability*. New York: Academic, 1973.
- [11] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*. New York: Academic, 1975.
- [12] M. G. Safonov, E. A. Jonckheere, and D. J. N. Limebeer, "Synthesis of positive real multivariable systems," *Int. J. Contr.*, vol. 45, pp. 817-842, 1987.
- [13] S. Boyd and Q. Yang, "Structured and simultaneous Lyapunov functions for system stability problems," *Int. J. Contr.*, vol. 49, pp. 2215-2240, 1989.
- [14] W. M. Haddad and D. S. Bernstein, "Robust stabilization with positive real uncertainty: Beyond the small gain theorem," *Syst. Contr. Lett.*, vol. 17, pp. 191-208, 1991.
- [15] —, "Explicit construction of quadratic Lyapunov functions for the small gain, positivity, circle, and popov theorems and their application to robust stability, part I: Continuous-time theory," *Int. J. Robust and Nonlinear Contr.*, vol. 3, pp. 313-339, 1993.
- [16] R. J. Benhabib, R. P. Iwens, and R. L. Jackson, "Stability of large space structure control systems using positivity concepts," *AIAA J. Guid. Contr.*, vol. 4, pp. 487-494, 1981.
- [17] S. M. Joshi, *Control of Large Flexible Space Structures*. New York: Springer-Verlag, 1989.
- [18] J. J. Bongiorno, Jr., "Real-frequency stability criteria for linear time-varying systems," *Proc. IEEE*, vol. 52, pp. 832-841, 1964.
- [19] I. W. Sandberg, "A frequency-domain condition for the stability of feedback systems containing a single time-varying nonlinear element," *Bell Sys. Tech. J.*, vol. 43, pp. 1601-1638, 1964.
- [20] G. Zames, "On the input-output stability of time-varying nonlinear feedback systems, part II: Conditions involving circles in the frequency plane and sector nonlinearities," *IEEE Trans. Automat. Contr.*, vol. AC-11, pp. 465-476, 1966.
- [21] J. L. Willems, "The circle criterion and quadratic Lyapunov functions for stability analysis," *IEEE Trans. Automat. Contr.*, vol. AC-18, p. 184, 1973.
- [22] P. Molander and J. C. Willems, "Synthesis of state feedback control laws with a specified gain and phase margin," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 928-931, 1980.
- [23] P. P. Khargonekar, I. R. Petersen, and K. Zhou, "Robust stabilization of uncertain linear systems: Quadratic stability and H^∞ control theory," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 356-361, 1990.
- [24] B. R. Barmish and C. L. DeMarco, "A new method for improvement of robustness bounds for linear state equations," in *Proc. Conf. Infor. Sci. Sys.*, Princeton, NJ, 1986, pp. 115-120.
- [25] M. A. Leal and J. S. Gibson, "A first-order Lyapunov robustness method for linear systems with uncertain parameters," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 1068-1070, 1990.
- [26] V. M. Popov, "Absolute stability of nonlinear systems of automatic control," *Automation and Remote Contr.*, vol. 22, pp. 857-875, 1962.
- [27] K. S. Narendra and J. H. Taylor, "Stability of nonlinear time-varying systems," *IEEE Trans. Automat. Contr.*, vol. AC-12, pp. 628-629, 1967.
- [28] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*. New York: Springer-Verlag, 1979.
- [29] D. S. Bernstein and W. M. Haddad, "Robust stability and performance analysis for state space systems via quadratic Lyapunov bounds," *SIAM J. Matrix Anal. Appl.*, vol. 11, pp. 239-271, 1990.
- [30] —, "Parameter-dependent Lyapunov functions and the discrete-time Popov criterion for robust analysis," *Automatica*, vol. 30, pp. 1015-1021, 1994.
- [31] —, "The parabola test: A unified extension of the circle and Popov criteria," in *Proc. Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 2662-2663.
- [32] D. C. Hyland and D. S. Bernstein, "The optimal projection equations for fixed-order dynamic compensation," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 1034-1037, 1984.
- [33] D. S. Bernstein and W. M. Haddad, "LQG Control with an H_∞ performance bound: A Riccati equation approach," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 293-305, 1989.

Design of L-Q Regulators for State Constrained Continuous-Time Systems

Carlos E. T. Dórea and Basilio E. A. Milani

Abstract—A new methodology to the design of L-Q regulators for continuous-time systems subject to linear state constraints is proposed, consisting of two parts. In the first one, the positive invariance of the polyhedron defined by the constraints in the state space is imposed, guaranteeing thereby that the constraints will not be violated. Furthermore, the admissible constrained controllers are parameterized via the determination of the elements which are fixed in the state feedback matrix. This parameterization enables in a second part the L-Q regulator to be obtained from the solution of a parameter optimization problem subject to linear constraints, for which it is proposed a specialized feasible directions method.

I. INTRODUCTION

In most of engineering applications, linear systems subject to state and input constraints are frequently found, since such constraints are generally associated to physical limitations in the course of the variables or to the validation domain of the linearization of nonlinear models. The controllers to be designed for such systems should therefore be capable to attain their performance specifications without violating the constraints.

Due to its known attractive properties, a classical control systems design specification which is widely used is the optimal linear-quadratic (L-Q) regulation, which aims to achieve a compromise between the speed of convergence to the zero state and the magnitude of the input amplitudes necessary to that, by minimizing a quadratic performance index [1]. The problem of L-Q regulation of systems under constraints has been treated to present in the framework of optimal control theory by means of variational calculus techniques, of difficult application because complex, generally resulting, moreover, in an open-loop "bang-bang" control law (see e.g., [2]). Other techniques are based on the choice of weighting matrices for which the solution of the unconstrained optimal L-Q regulator problem results in a control law that respects the constraints [3], [4]. Such techniques imply clearly in a limitation to the designer, since the weighting matrices cannot be freely chosen in order to satisfy performance requirements.

The forthcoming of new techniques to the design of controllers for linear constrained systems has been propelled lately by the development of the theory of positively invariant polyhedra, that are domains from which the state vector trajectories cannot escape (see e.g., [5], [6] and references therein). The basic idea behind the techniques derived from this theory is to determine a control law that imposes the positive invariance of the polyhedron defined in the state space by the linear constraints on the state vector. In order to solve this problem, Castelan and Hennet [7], by interpreting geometrically the positive invariance conditions, have proposed elaborate eigenstructure assignment techniques.

The aim of this paper is to develop a new methodology to the design of L-Q regulators for continuous-time systems under linear state constraints, applying to this end results from the invariant polyhedra theory and parameter optimization techniques. This objective

is achieved in two stages. In the first one, by an appropriate change in the state-space representation of the system and the utilization of results from the geometric theory of control systems [8], the conditions for positive invariance are transformed into a set of linear constraints on the elements of state feedback matrices, which are thereby parameterized. As a consequence of this parameterization, the problem of positive invariance of an unbounded polyhedron with simultaneous closed-loop stability is split into two distinct ones for lower order systems: a stabilization problem and a problem of positive invariance of a bounded polyhedron. In the second stage, the original problem of state constrained L-Q regulation is reformulated as a parameter optimization problem subject to the constraints defined in the first stage. To solve this problem, it is proposed a specialized feasible directions method which is initialized with a point obtained from the solution of an auxiliary linear programming problem.

This paper is organized as follows. Section II contains preliminaries. In Section III the positive invariance of the polyhedron defined by the constraints is achieved together with the parameterization of the controllers. Section IV presents the design of the constrained L-Q regulators. An example in Section V illustrates the application of the proposed design and some conclusions are found in Section VI.

II. PRELIMINARIES

Consider the linear, time-invariant, continuous-time system

$$\dot{x}(t) = Ax(t) + Bu(t) \quad (1)$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^m$ is the input vector and the system is supposed to be stabilizable. Under a state feedback control law, $u(t) = Fx(t)$, it can be written as

$$\dot{x}(t) = (A + BF)x(t). \quad (2)$$

Definition 1 [5]: A nonempty set Ω is a positively invariant set of system (2) if and only if for any initial state $x(0) \in \Omega$ the state vector trajectory remains in Ω .

Assume now that system (2) is subject to symmetrical linear constraints, $[-\rho \leq Gx(t) \leq \rho]$, where $G \in \mathbb{R}^{r \times n}$, $r \leq n$ is a full rank matrix and the elements ρ_i of vector $\rho \in \mathbb{R}^r$ are such that $\rho_i > 0$. These constraints define in the state space the convex symmetrical polyhedron

$$S[G, \rho] = \{x \in \mathbb{R}^n / -\rho \leq Gx \leq \rho\} \quad (3)$$

where the inequalities are componentwise. For $r < n$, $S[G, \rho]$ is clearly unbounded.

Proposition 1: The symmetrical polyhedron (see also [6]) $S[G, \rho]$ is a positively invariant set of system (2) if and only if there exist matrices H and $K \in \mathbb{R}^{r \times n}$, with $H_{ij} \geq 0$ for $i \neq j$ and $K_{ij} \geq 0$ for any i, j such that

$$(H - K)G = G(A + BF) \quad (4)$$

$$(H + K)\rho \leq 0. \quad (5)$$

Proof: Let $R[T, \omega]$ be a general, not necessarily symmetrical convex polyhedron defined by

$$R[T, \omega] = \{x \in \mathbb{R}^n / Tx \leq \omega\} \quad (6)$$

Manuscript received October 4, 1993; revised April 21, 1994.

C. E. T. Dórea is with LAAS/CNRS, 7, Avenue du Colonel Roche, 31077 Toulouse, France.

B. E. A. Milani is with DT, Faculty of Electrical Engineering, UNICAMP, CP 6101, 13081-970 Campinas, SP, Brazil.

IEEE Log Number 9407227.

where $T \in \mathbb{R}^{n \times n}$ and $\omega \in \mathbb{R}^n$, $\omega_i > 0$. In [5] and [6] it is shown that $R[T, \omega]$ is positively invariant for system (2) if and only if there exists a matrix $J \in \mathbb{R}^{n \times n}$, with $J_{ij} \geq 0$ for $i \neq j$ such that

$$JT = T(I + BF) \quad (7)$$

$$J\omega \leq 0 \quad (8)$$

The polyhedron $S[G, \rho]$ can be written as a generic one (6). One only needs to consider

$$T = \begin{bmatrix} G \\ -G \end{bmatrix} \quad \omega = \begin{bmatrix} \rho \\ \rho \end{bmatrix} \quad (9)$$

$$J = \begin{bmatrix} H & K \\ I & M \end{bmatrix} \quad (10)$$

with $H_{ij}, M_{ij} \geq 0$ for $i \neq j$ and $K_{ij}, I_{ij} \geq 0$ for any i, j . Substituting (9)–(10) in (7)–(8) one arrives to (4)–(5) and two other redundant equations. \square

III POSITIVE INVARIANCE OF $S[G, \rho]$

This section is concerned with the obtention of a proportional state feedback law $u = Fx$ for which the symmetrical polyhedron $S[G, \rho]$ is positively invariant and the closed loop system (2) is asymptotically stable. Consider then the following change of basis in the system (1)

$$x = Qz \quad (11)$$

where $Q \in \mathbb{R}^{n \times n}$ is an orthogonal matrix such that

$$G_+ - G_-Q = \begin{bmatrix} 0 & G_- \end{bmatrix} \quad G_- \in \mathbb{R}^{n \times n} \quad (12)$$

The new representations of system (1) and polyhedron $S[G, \rho]$ are

$$\dot{z} = A_1 z + B_1 u \quad (13)$$

$$A = Q^{-1}AQ = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad A \in \mathbb{R}^{n \times n} \quad (14)$$

$$B = Q^{-1}B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \quad B \in \mathbb{R}^{n \times m} \quad (15)$$

$$S[G, \rho] = \{z \in \mathbb{R}^n / -\rho \leq G_+ z \leq \rho\} \quad (16)$$

Proposition 2 The polyhedron $S[G, \rho]$ is positively invariant for system (2) if and only if

- 1) There exists a matrix $F_1 \in \mathbb{R}^{m \times n}$ such that

$$A_{11} + B_1 F_1 = 0 \quad (17)$$

- 2) There exist matrices $F_2 \in \mathbb{R}^{m \times n}$, $H \in \mathbb{R}^{n \times n}$, with $H_{ij} \geq 0$ for $i \neq j$ and $K_{ij} \geq 0$ for any i, j such that

$$(H - K)G_2 = G_2(A_{22} + B_2 F_2) \quad (18)$$

$$(H + K)\rho \leq 0 \quad (19)$$

Proof Consider the state feedback matrix in the transformed system (13)

$$F = FQ = [F_1 \ F_2] \quad F_2 \in \mathbb{R}^{m \times n} \quad (20)$$

Its substitution together with G_+ (12), A (14) and B (15) in (4) yields

$$[0 \ (H - K)G_-] = [G_2(A_{21} + B_2 F_1) \ G_2(A_{22} + B_2 F_2)] \quad (21)$$

To satisfy this equation it is necessary that $G_2(A_{21} + B_2 F_1) = 0$. Since G_+ (12) has linearly independent columns, this condition holds if and only if (17) holds. Condition 2 may be directly verified from (5) and (21). \square

Considering (17) satisfied, it can be noted from (18)–(19) that the original problem of positive invariance of $S[G, \rho]$ for the system represented by the pair (A, B) has been reduced to the positive invariance of the bounded polyhedron $S[G_+, \rho]$ for a lower order system represented by the pair (A_{22}, B_2) . This reduction becomes significative mainly for large dimensions systems subject to few constraints, since redundant equalities are eliminated.

It is easy to verify that condition (17) is equivalent to the (A, B) -invariance of $\ker(G_-)$, the null space of the matrix G_- , result firstly obtained by Castelan and Hennet [7]. This interpretation enabled them to apply results from the so called geometric approach to control systems [8] in the analysis of stabilizability and controllability of state constrained systems, and to reformulate the design problem as an eigenstructure assignment problem.

In this paper, geometric theory results will be applied in a quite different manner. Suppose that (17) holds. In closed loop one has

$$A_1 = A + BF = \begin{bmatrix} A_{11} + B_1 F_1 & A_{12} + B_1 F_2 \\ 0 & A_{22} + B_2 F_2 \end{bmatrix} \quad (22)$$

The eigenvalues of A_1 are therefore those of submatrices $A_{11} = A_{11} + B_1 F_1$ and $A_2 = A_{22} + B_2 F_2$. The positive invariance of $S[G, \rho]$ implies certain spectral properties for A_{11} [5]–[6]. In particular, if the inequalities (19) are strictly satisfied, A_1 is asymptotically stable. The problem of positive invariance of $S[G, \rho]$ with closed loop stability can therefore be summarized as follows:

- Find F_1 such that (17) holds and the eigenvalues of $A_{11} + B_1 F_1$ have negative real parts.
- Find F_2 such that (18) holds and (19) holds with strict inequalities.

The stabilization of A_{11} , which represents the map of A_F restricted to $\ker(G_-)$, is dependent upon the maximal controllability subspace contained in $\ker(G_-)$ [8], here denoted as R^* . This subspace defines what can be controlled in $\ker(G_-)$ by matrices F that satisfy (17). An efficient and numerically stable method for its computation has been presented in [9] (see also [10]). It consists in finding orthogonal transformation matrices which change the coordinate basis of the system in such a way that the structure of the desired subspace is made evident. In particular, it is always possible to compute changes of basis [9]

$$u = Zu \quad z = Q_R z \quad (23)$$

such that the system (13) is represented as follows

$$\dot{z} = A_1 z + B_1 u \quad (24)$$

$$\hat{A} = Q'_R \hat{A} Q_R = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} & \hat{A}_{13} \\ \hat{A}_{21} & \hat{A}_{22} & \hat{A}_{23} \\ \hat{A}_{31} & \hat{A}_{32} & \hat{A}_{33} \end{bmatrix}; \quad \hat{A}_{11} \in \mathbb{R}^{c \times c} \quad (25)$$

$$\hat{B} = Q'_R \hat{B} Z = \begin{bmatrix} \hat{B}_{11} & \hat{B}_{12} \\ \hat{B}_{21} & 0 \\ \hat{B}_{31} & 0 \end{bmatrix}; \quad \hat{B}_{11} \in \mathbb{R}^{c \times b} \quad (26)$$

where $\hat{B}_{31} \in \mathbb{R}^{r \times b}$ has full column rank

$$\hat{F} = Z' \hat{F} Q_R = \begin{bmatrix} \hat{F}_{11} & \hat{F}_{12} & \hat{F}_{13} \\ \hat{F}_{21} & \hat{F}_{22} & \hat{F}_{23} \end{bmatrix}; \quad \hat{F}_{11} \in \mathbb{R}^{b \times c} \quad (27)$$

$$\hat{G} = \hat{G} Q_R = [0 \quad 0 \quad \hat{G}_3]; \quad \hat{G}_3 \in \mathbb{R}^{r \times r} \quad (28)$$

$$\hat{\mathcal{R}}^* = \text{Im} \begin{bmatrix} I_c \\ 0 \\ 0 \end{bmatrix} \quad \ker(\hat{G}) = \begin{bmatrix} I_c & 0 \\ 0 & I_{n-(c+r)} \\ 0 & 0 \end{bmatrix} \quad (29)$$

where I_n denotes the identity matrix of order n

$$\hat{S}[\hat{G}, \rho] = \{\hat{x} \in \mathbb{R}^n / -\rho \leq \hat{G}\hat{x} \leq \rho\}. \quad (30)$$

Let \mathbf{F} be the set of matrices F for which $\ker(G)$ is (A, B) -invariant. Since \hat{B}_{31} has linearly independent columns, the set $\hat{\mathbf{F}}$ is uniquely defined by

$$\hat{\mathbf{F}} \equiv \left\{ \hat{F} / [\hat{A}_{31} \quad \hat{A}_{32}] + \hat{B}_{31} [\hat{F}_{11} \quad \hat{F}_{12}] = 0 \right\} \quad (31)$$

and in closed loop, for $\hat{F} \in \hat{\mathbf{F}}$

$$\hat{A}_F = \hat{A} + \hat{B}\hat{F} = \begin{bmatrix} \hat{A}_{F11} & \hat{A}_{F12} & \hat{A}_{F13} \\ 0 & \hat{A}_{F22} & \hat{A}_{F23} \\ 0 & 0 & \hat{A}_{F33} \end{bmatrix} \quad (32)$$

where the eigenvalues of \hat{A}_{F11} may be freely assigned while those of \hat{A}_{F22} cannot be modified by any $\hat{F} \in \hat{\mathbf{F}}$. If one considers the output equation $y(t) = Gx(t)$, the eigenvalues of \hat{A}_{F22} are the so-called transmission zeros (invariant zeros for some authors) [8], [10]. Therefore, if any of them has positive or null real part, the positive invariance of $S[G, \rho]$ with simultaneous closed-loop asymptotical stability cannot be achieved. In this case, a possible solution is to add adequately constraints to the states so as to assign only the eigenvectors corresponding to the stable eigenvalues of \hat{A}_{F22} to the null space of the new matrix G [7], [11].

Assuming henceforth that the eigenvalues of \hat{A}_{F22} have negative real parts, it can be observed from the representation (24)-(31) that the original problem of positive invariance of $S[G, \rho]$ with stability can be split into two distinct problems as follows:

Problem 1: Given \hat{F}_{11} and \hat{F}_{12} fixed by (31), find \hat{F}_{21} such that the eigenvalues of $\hat{A}_{11} + \hat{B}_{11}\hat{F}_{11} + \hat{B}_{12}\hat{F}_{21}$ have negative real parts.

Problem 2: Find $\hat{F}_{13} \in \mathbb{R}^{b \times r}$ such that there exist H and K which satisfy

$$(H - K)\hat{G}_3 = \hat{G}_3(\hat{A}_{33} + \hat{B}_{31}\hat{F}_{13}) \quad (33)$$

$$(H + K)\rho < 0 \quad (34)$$

with $H_{ij} \geq 0$ for $i \neq j$ and $K_{ij} \geq 0$ for any i, j .

Any stabilization procedure may be used to solve Problem 1. Regarding Problem 2, a simple and efficient approach to handle it is to formulate a linear program (LP) having conditions (33), (34) as constraints. A scheme similar to that proposed in [12], where the objective function is chosen so as to increase the rate of convergence of the state to the equilibrium, cannot be applied here because there are no bounds on the gains of \hat{F}_{13} . This difficulty can be removed by choosing as objective function to be minimized, for instance, the sum of the absolute values of the elements of \hat{F}_{13} . The solution of Problem 2 can therefore be obtained from the following LP

$$\min_{\hat{F}_{13}, H, K, W} \sum_{i=1}^b \sum_{j=1}^r (\hat{F}_{13,ij} + 2W_{ij}) \quad (35)$$

sub.to: $(H - K)\hat{G}_3 = \hat{G}_3(\hat{A}_{33} + \hat{B}_{31}\hat{F}_{13})$

$$(H + K)\rho \leq -\sigma\rho$$

$$\hat{F}_{13} + W \geq 0$$

$$H_{ij} \geq 0 \text{ for } i \neq j; K_{ij} \geq 0; W_{ij} \geq 0.$$

In view of the constraints involved, it can be easily verified that the minimization of the objective function yields $W_{ij} = 0$ for $\hat{F}_{13,ij} \geq 0$ and $W_{ij} = -\hat{F}_{13,ij}$ for $\hat{F}_{13,ij} \leq 0$. Therefore, at the optimum, $\hat{F}_{13,ij} + 2W_{ij} = |\hat{F}_{13,ij}|$.

The small positive number σ must be chosen so as to guarantee the strict satisfaction of inequalities (34) as well as to avoid an excessive proximity to the origin of the eigenvalues of $(\hat{A}_{33} + \hat{B}_{31}\hat{F}_{13})$. In fact, it can be easily shown [12] that $-\sigma$ defines an upper bound on the real part of such eigenvalues.

Before continuing, it should be remarked that the preceding development has been motivated by the case $r = \text{rank}(G) < n$. When $r = n$ the polyhedron $S[G, \rho]$ is bounded and the changes of basis (11), (23) are not required. In this case, a stabilizing F for which $S[G, \rho]$ is positively invariant may be directly obtained from the resolution of (35) using the corresponding matrices of system (2) and constraints (3), in their original basis. The same is applicable to the L-Q regulator design, to be treated in the next section.

IV. CONSTRAINED L-Q REGULATOR DESIGN

The resolution of Problems 1 and 2 defined in last section provides, if possible, a controller such that $S[G, \rho]$ is positively invariant and the closed-loop system is asymptotically stable. In a typical control system design, however, performance specifications are also included and among them the optimal L-Q regulation is widely used. Consider then the quadratic performance index

$$J(u(t)) = \mathbf{E} \left\{ \int_0^\infty [x(t)' S_1' x(t) + u(t)' S_2 u(t)] dt \right\} \quad (36)$$

where \mathbf{E} is the mathematical expectation operator; S_1 and S_2 are constant real matrices; S_2 is positive definite; $x_0 = x(0)$ is a random vector with $\mathbf{E}\{x_0\} = 0$ and $\mathbf{E}\{x_0 x_0'\} = X_0' X_0$.

The objective of the state constrained L-Q regulator problem to be solved here is to find a state feedback law $u = Fx$ such that $S[G, \rho]$ is positively invariant and the performance index (36) is simultaneously minimized. Consider then the changes of basis (11), (23) which lead the system to the form (24)-(31). Considering also the state feedback law $\hat{u} = \hat{F}\hat{x}$, the index (36) may be rewritten as

$$J(\hat{F}) = \mathbf{E} \left\{ \int_0^\infty \hat{x}(t)' [\hat{S}_1' \hat{S}_1 + \hat{F}' \hat{S}_2 \hat{F}] \hat{x}(t) dt \right\} \quad (37)$$

where $S_1 = S_1(QQ_R)$, $S_2 = Z'S_2Z$, $E\{x_0\} = 0$ and $E\{x_0x_0'\} = (QQ_R)X_0'X_0(QQ_R)' = X_0'X_0$. The performance index $J(F)$ is given by [13]

$$J(F) = T_1(\lambda_0' \lambda_0 I) \quad (38)$$

$$(\hat{A} + \hat{B}F)'T + T(\hat{A} + BF) + S_1'S_1 + F'S_2F = 0 \quad (39)$$

We are now in position to formulate the state constrained LQ regulator design problem

$$\begin{aligned} \min_{F \in \mathbf{F}} J(F) \\ \text{subject to } (H - K)G_1 &= G_1(A_{11} + B_{11}F_{11}) \\ (H + K)\rho &\leq 0 \\ F \in \mathbf{F}, H_{ij} &\geq 0 \text{ for } i \neq j, K_{ij} \geq 0 \end{aligned} \quad (40)$$

Matrices $F \in \mathbf{F}$ (31) have fixed and free elements. Let $\underline{\alpha}$ be a vector composed by the free elements of $F \in \mathbf{F}$ together with the elements of H and K . The problem (41) can therefore be rewritten as the following standard linearly constrained minimization problem

$$\min J(\underline{\alpha}) \quad (41)$$

$$\text{subject to } C\underline{\alpha} \leq \underline{f} \quad (42)$$

Without considering the constraints (42) $\min J(\underline{\alpha})$ (41) corresponds to a parameter optimization problem in LQ regulators subject to structural constraints. For solution of this kind of problem Dorca and Milani [13] proposed a specialized hybrid descent optimization procedure composed by a sequence of steps of Modified Newton, Newton's, and Quasi Newton methods. For obtaining a good trade off between the cost per step and the convergence rate the methods are chosen along the descent process considering their expected computational effort per step, an evaluation of the convergence rate of the descent sequence and the proximity to the optimal point.

To solve the parameter optimization problem with linear constraints (41) (42) the method proposed in [13] is transformed into a feasible directions optimization method [14] described in the sequel.

Assuming

$\nabla J(\underline{\alpha})$ gradient vector of $J(\underline{\alpha})$

$H(\underline{\alpha})$ hessian matrix of $J(\underline{\alpha})$

$H_+(\underline{\alpha})$ positive definite approximation to $H(\underline{\alpha})$

$H_d(\underline{\alpha})$ matrix obtained from the diagonal scaling of $H(\underline{\alpha})$ fulfilling the secant condition of Quasi Newton methods,

as defined in [13], the following active set based procedure [14] is proposed to solve the constrained minimization problem (41), (42), starting from a feasible point $\underline{\alpha}_0$

1) Initialization

- Define the tolerance ϵ
- Set

$$l = 0 \quad \underline{\alpha}_l = \underline{\alpha}_0$$

2) Active set computations

- Construct C'_A , the active constraints matrix composed by the lines C'_i of C' (42) such that

$$C'_i \underline{\alpha}_l = f_i$$

3) Computation of the projection matrix

- Compute the orthogonal matrix V by performing the following QR factorization [14]

$$C'_A = VU = [V_1 \quad V_2] \begin{bmatrix} U_1 \\ 0 \end{bmatrix}$$

where the matrix U_1 has full row rank

4) Test for convergence

- Compute the reduced gradient vector

$$q(\underline{\alpha}_l) = V_2' \nabla J(\underline{\alpha}_l)$$

- If $\|q(\underline{\alpha}_l)\| > \epsilon$ go to step 5. Otherwise compute the Lagrange multipliers vector λ by solving the system of equations

$$U_1 \lambda = V_1' \nabla J(\underline{\alpha}_l)$$

- If $\lambda \geq 0$ STOP and take $\underline{\alpha}_l$ as the solution. Otherwise eliminate the row of C'_A correspondent to the most negative component of λ and return to step 3.

5) Computation of a feasible search direction

- Based on the directions proposed in [13] select the descent method and define a positive definite matrix $L(\underline{\alpha}_l)$ considering the following alternatives

$$L(\underline{\alpha}_l) = \begin{cases} H(\underline{\alpha}_l) & \text{modified Newton} \\ H(\underline{\alpha}_l) & \text{Newton's} \\ H_d(\underline{\alpha}_l) & \text{Quasi Newton} \end{cases}$$

- Compute

$$M(\underline{\alpha}_l) = V_2 L(\underline{\alpha}_l) V_2'$$

- Solve the linear system of equations

$$M(\underline{\alpha}_l) p_l = -q(\underline{\alpha}_l)$$

- Compute

$$\underline{d}_l = V_2 p_l$$

6) Solution update

- Find a step size γ such that

$$\underline{\alpha}_{l+1} = \underline{\alpha}_l - \gamma \underline{d}_l \text{ is feasible}$$

$$J(\underline{\alpha}_{l+1}) < J(\underline{\alpha}_l)$$

- Set

$$l = l + 1$$

- Return to step 2

Remarks

- 1) The existence of γ in step 6 is assured by the positive definiteness of matrices $L(\underline{\alpha}_l)$, $M(\underline{\alpha}_l)$, in step 5 [15]. Due to the non convexity of $J(\underline{\alpha})$ it is not possible to assure that the attained minimum point in step 4 is the global one. To the author's knowledge, a convex programming approach to problems with the structural constraints present in (41), is not available to date.

- 2) The optimal state feedback matrix F^* in the original basis (2) is obtained from the optimal solution $F(\underline{\alpha}^*)$ by

$$F^* = ZF(\underline{\alpha}^*)(QQ_R)'$$

V. NUMERICAL EXAMPLE

The following example has been borrowed from [5]. Consider the system for which

$$A = \begin{bmatrix} 9.10 & 0.47 & -6.33 \\ 7.62 & 0 & 7.56 \\ 2.62 & -3.28 & 9.91 \end{bmatrix} \quad B = \begin{bmatrix} 1.82 & 3.61 \\ 1.24 & -3.77 \\ -4.91 & 0 \end{bmatrix}.$$

The state constraints are defined by

$$G = \begin{bmatrix} 5.69 & 1.97 & -1.68 \\ 2.24 & -1.68 & 5.59 \end{bmatrix} \quad \rho = \begin{bmatrix} 8.75 \\ 10.50 \end{bmatrix}.$$

The system is transformed into the representation (24)-(31) yielding

$$\hat{A} = \begin{bmatrix} 1.7936 & 1.8674 & 12.9823 \\ & 11.1206 & -2.1901 \\ -1.6275 & 1.1278 & 6.0957 \end{bmatrix}$$

$$\hat{B} = \begin{bmatrix} -1.1843 & -4.1421 \\ & 3.4836 \\ 4.6808 & 0 \end{bmatrix} \quad \hat{G} = \begin{bmatrix} 0 & 6.2514 & 0.0017 \\ 0 & 0.0055 & 6.2520 \end{bmatrix}.$$

The matrices $\hat{F} \in \hat{\mathcal{F}}$ are given by

$$\hat{F} = \begin{bmatrix} 0.3477 & X & X \\ 1.0678 & X & X \end{bmatrix}$$

where the elements X are free to assume any value. For such matrices, it can be readily verified that the closed-loop system will have -3.0399 as a fixed eigenvalue.

Using $S_1 = I_3$; $S_2 = I_2$; $X_0'X_0 = \frac{1}{3}I_3$ and starting from the solution of the LP (34) with $\sigma = 0.1$, the resolution of the parameter optimization problem (42) yields

$$J(\hat{F}^*) = 4.1255$$

$$\hat{F}^* = \begin{bmatrix} 0.3477 & -2.5595 & -4.6649 \\ 1.0678 & -7.4796 & -1.9137 \end{bmatrix}.$$

In the original basis,

$$F^* = \begin{bmatrix} -0.2561 & -1.0943 & 3.2597 \\ -8.7191 & -0.3818 & -1.0603 \end{bmatrix}.$$

The optimal solution of the unconstrained L-Q regulator problem is $J^* = 1.1228$. However, for $x(0) = [0.6161 \ 5.4503 \ 3.2695]'$, it can be seen in Fig. 1 that the state vector trajectory resulting from this solution violates the imposed constraints.

VI. CONCLUSION

In this paper a new approach has been proposed to the design of L-Q regulators for state constrained continuous-time systems. The design problem has been stated as a parameter optimization problem subject to linear constraints. The respect to the constraints has been guaranteed by achieving closed-loop positive invariance of the polyhedron defined by them. By appropriately changing the state-space representation of the system, the positive invariance of an unbounded polyhedron with simultaneous closed-loop stability has been achieved by solving two distinct problems for lower order systems: a stabilization problem and a problem of positive invariance of a bounded polyhedron, which can be solved by properly defining a linear programming problem. Moreover, the admissible constrained controllers have been parameterized by means of the determination of the fixed elements in the state feedback matrix. The conditions

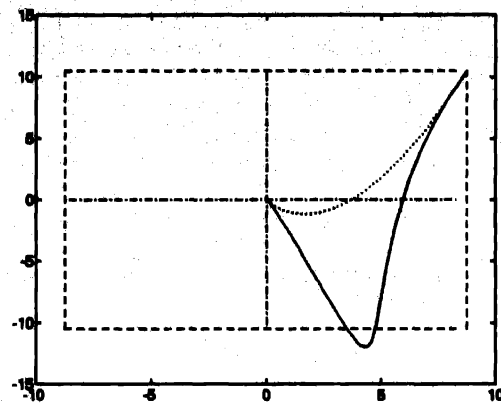


Fig. 1. Trajectories in projection $(Gx)_1 \times (Gx)_2$ for the constrained (..) and unconstrained (-) L-Q regulator.

for existence of such controllers are directly verified from structural properties of the matrices of the system at each step of the design. A specialized feasible directions method has been proposed to solve the resulting constrained parameter optimization problem. The proposed design makes no assumptions, other than the usual ones, on the weighting matrices and is very simple when compared to other approaches available in the literature.

REFERENCES

- [1] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*. New York: Wiley, 1972.
- [2] D. D. Thompson and R. A. Volz, "The linear-quadratic cost problem with linear state constraints and the nonsymmetric Riccati equation," *SIAM J. Contr.*, vol. 13, pp. 110-145, 1975.
- [3] R. Castelein and A. Johnson, "Constrained optimal control," *IEEE Trans. Automat. Contr.*, vol. AC-34, no. 1, pp. 122-126, 1989.
- [4] D. Mehdi, M. Zasadzinski, and M. Darouach, "Design of constrained optimal feedback law from the viewpoint of the inverse optimal regulator," in *Proc. 1993 American Contr. Conf.*, San Francisco, CA, June 1993, pp. 3185-3186.
- [5] E. B. Castelan and J. C. Hennet, "On invariant polyhedra of continuous-time linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-38, pp. 1680-1685, 1993.
- [6] G. Bitsoris, "Existence of polyhedral positively invariant sets for continuous-time systems," in *Control—Theory and Advanced Technology*, vol. 7, pp. 407-427, 1991.
- [7] E. B. Castelan and J. C. Hennet, "Eigenstructure assignment for state constrained linear continuous time systems," *Automatica*, vol. 28, no. 3, pp. 605-611, 1992.
- [8] W. M. Wonham, *Linear Multivariable Control. A Geometric Approach*. New York: Springer-Verlag, 1979.
- [9] A. Linnemann, "A condensed form for disturbance decoupling with simultaneous pole placement using state feedback," in *Proc. 10th IFAC World Congress*. Munich, vol. 9, 1987, pp. 92-9.
- [10] P. M. Van Dooren, "The generalized eigenstructure problem in linear system theory," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 111-129, 1981.
- [11] J. C. Hennet and E. B. Castelan, "constrained control of unstable multivariable linear systems," in *Proc. of 1993 European Contr. Conf.*, vol. 4, 1993, pp. 2039-2043.
- [12] M. Vassilaki and G. Bitsoris, "constrained regulation of linear continuous-time dynamical systems," *Syst. Contr. Lett.*, vol. 13, pp. 247-252, 1989.
- [13] C. E. T. Dórea and B. E. A. Milani, "A computational method for optimal L-Q regulation with simultaneous disturbance decoupling," *Proc. of 1993 American Contr. Conf.*, San Francisco, CA, June 1993, pp. 2669-2672. Also accepted for *Automatica*.
- [14] P. E. Gill, W. Murray and M. H. Wright, *Practical Optimization*. London: Academic, 1981.
- [15] D. G. Luenberger, *Linear and Non-Linear Programming*. Redwood City, CA: Addison-Wesley, 1984.

The Finite Inclusions Theorem

Richard D. Kaminsky and Theodore E. Djaferis

Abstract—This paper presents a novel, necessary and sufficient condition for a polynomial to have all its roots in an arbitrary convex region of the complex plane. The condition may be described as a variant of Nyquist's stability theorem; however, unlike this theorem it only requires knowledge of the polynomial's value at finitely many points along the region's boundary. A useful corollary, the Finite Inclusions Theorem (FIT), provides a simple sufficient condition for a family of polynomials to have its roots in a given convex region. Since FIT only requires knowledge of the family's value set at finitely many points along the region's boundary, this corollary provides a new convenient tool for the analysis and synthesis of robust controllers for parametrically uncertain systems.

1 INTRODUCTION

With regards to linear time invariant (LTI) systems there are now two well established ways to model system uncertainty: either parametrically or through H_∞ norm bounded classes of transfer functions. In the former case, which we are mainly interested in here, both robust stability and performance questions often reduce to showing that a certain parametric family of polynomials (or matrices) has its zeros (eigenvalues) in a given region of the complex plane. As engineers, we would like to have efficient numerical methods for verifying this inclusion. While many good methods and interesting results have come to light, our optimism in seeking a general solution to this problem, especially one that will permit us to conveniently synthesize robust controllers, must be tempered by the fact that there are certain fundamental mathematical difficulties in relating a polynomial's coefficients to its zeros.

Many if not all of the major results connecting a polynomial's coefficients with its zeros can be derived from the argument principle (aka Nyquist's theorem). Among them, the fact that a degree n polynomial has precisely n zeros (counting multiplicities), the algebraic root clustering criteria of Routh, Hurwitz, Schur, Cohn, and Jury, the Hermite-Biehler theorem, and Kharitonov's theorem. The purpose of this paper is to present a novel variant of the argument principle which is useful in solving robust parametric control problems.

Recall that Nyquist's theorem (i.e., the argument principle) states the following: Let $D \subset \mathbb{C}$ be an open, simply connected set, $f: D \rightarrow \mathbb{C}$ be a function analytic on D except for finitely many poles, and $\Gamma \subset D$ be a counterclockwise, closed, rectifiable, Jordan curve on which f has neither poles nor zeros. Further, let λ and λ_p be the number of zeros and poles of f inside Γ , and let λ be the net number of counterclockwise encirclements about the origin made by the curve $f(\Gamma)$. Then $\lambda - \lambda_p = \lambda$.

As a numerical method for determining pole/zero locations, Nyquist's theorem is flawed because it requires evaluating $f(\cdot)$ at the infinitely many points Γ . In practice, of course, we often choose finitely many points $\{s_k\} \subset \Gamma$ and from a plot of $\{f(s_k)\}$ make an educated guess as to the number of counterclockwise encirclements that $f(\Gamma)$ makes about the origin. Despite the room for error here, this procedure usually works quite well, yet it raises the question,

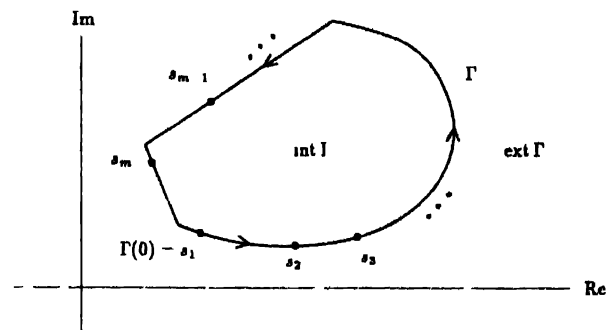


Fig. 1. A closed Jordan curve with convex interior.

Can λ ever be rigorously determined from some finite set of points $\{f(s_k)\}$?

The main result of this paper is that when f is a polynomial of degree $\leq n$ and the interior of Γ is convex, one can determine whether or not $\lambda = \lambda_p = n$ from an appropriately chosen set of points $\{f(s_k)\}$. This result, the finite Nyquist theorem (Theorem 1), leads naturally to a very useful sufficient condition for a family of polynomials to have all its zeros inside Γ , namely the finite inclusions theorem (Theorem 2).

Section II derives the finite Nyquist theorem and Section III derives the finite inclusions theorem. The case of Hurwitz stability was dealt with in [1]. For an application of these results to robust control analysis and synthesis problems, see [2]; the complete version of this paper.

II THE FINITE NYQUIST THEOREM

Let us introduce some notation. Let \mathbb{C} , \mathbb{R} , \mathbb{R}_+ , and \mathbb{Z} be the sets of complex, real, positive real, and integer numbers, respectively. Let $\arg \cdot$ be the argument of $\cdot \in \mathbb{C} \setminus \{0\}$ and $\text{Arg} \cdot$ the principle argument of $\cdot \in \mathbb{C} \setminus \{0\}$ with range $(-\pi, \pi]$. Let \cdot^* denote the complex conjugate of $\cdot \in \mathbb{C}$. Let $\text{ext } \Gamma$ be the exterior of a closed Jordan curve Γ , $\text{int } \Gamma$ the interior of a closed Jordan curve Γ , $\text{co } S$ the convex hull of S , and $\text{ext } Q$ the set of extreme points of a convex set Q .

Nyquist's theorem states that an n th degree polynomial p has all its zeros inside a counterclockwise closed Jordan curve Γ if and only if the point $p(s)$ revolves a net $2\pi n$ radians counterclockwise about the origin as s traverses Γ . An inconvenient feature of Nyquist's theorem is that $p(s)$ must be evaluated at the infinitely many points Γ . The following novel theorem states that when $\text{int } \Gamma$ is convex, $p(s)$ need only be evaluated at finitely many points. After this paper was accepted for publication, it was brought to our attention that a similar result was stated in [4].

Theorem 1—Finite Nyquist Theorem. Let $p(s) = \sum_{j=0}^n \alpha_j s^j$ where $n \geq 0$ and $\alpha_j \in \mathbb{C}$, and let $\Gamma \subset \mathbb{C}$ be a closed Jordan curve such that $\text{int } \Gamma$ is convex (see Fig. 1). Then, p is of degree n (i.e., $\alpha_n \neq 0$) and has all its zeros in $\text{int } \Gamma$ if and only if there exist $m \geq 1$ angles $\theta_k \in \mathbb{R}$ and a counterclockwise sequence of points $s_k \in \Gamma$, $1 \leq k \leq m$, such that

$$\forall 1 \leq k < m, |\theta_{k+1} - \theta_k| < \pi \quad (1a)$$

$$|2\pi n + \theta_1 - \theta_m| < \pi \quad (1b)$$

$$\forall 1 \leq k < m, p(s_k) \neq 0 \quad (1c)$$

$$\forall 1 \leq k < m, \arg p(s_k) \equiv \theta_k \pmod{2\pi} \quad (1d)$$

Manuscript received January 8, 1993; revised April 13, 1994.
R. D. Kaminsky is with Digital Storage Division, Shrewsbury, MA 01545, USA.

T. E. Djaferis is with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003-4410, USA.
IEEE Log Number 9407226.

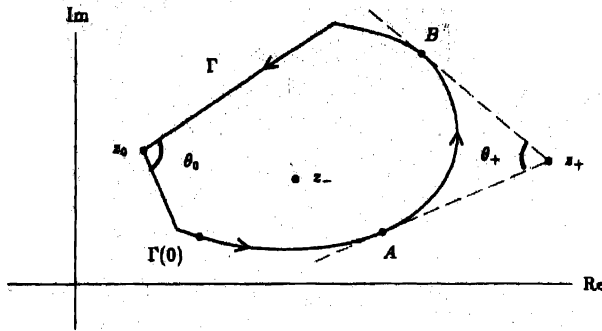


Fig. 2. A closed Jordan curve with convex interior.

Proof: The proof requires several preliminary definitions. First, let $I = [0, 1)$ and define the total variation [3, p. 328] of a function $f: I \rightarrow \mathbb{R}$ to be

$$Vf = \sup_{\substack{n \geq 1 \\ 0 \leq x_1 < x_2 < \dots < x_n < 1}} \sum_{j=1}^{n-1} |f(x_{j+1}) - f(x_j)|.$$

Intuitively, Vf is the total "vertical distance" through which f oscillates on I . Note, Vf is a nonnegative (possibly infinite) quantity, and by the elementary inequality $|\alpha + \beta| \leq |\alpha| + |\beta|$, $\alpha, \beta \in \mathbb{R}$, we have:

Lemma 1: For arbitrary $f, g: I \rightarrow \mathbb{R}$, $V(f + g) \leq Vf + Vg$.

Second, let us parameterize the closed Jordan curve Γ (see Fig. 1) by a continuous function $\Gamma: I \rightarrow \mathbb{C}$ where $\Gamma(0) = s_1$. For convenience both the curve and function will be denoted by Γ , and the left-hand limit $\lim_{t \rightarrow 1^-} \Gamma(t)$ will be denoted by $\Gamma(1^-)$. The above adjective closed refers to the property $\Gamma(0) = \Gamma(1^-)$ while the adjective Jordan refers to the property $\forall x \neq y, \Gamma(x) \neq \Gamma(y)$ (i.e., Γ does not intersect itself). Note, an intuitive—but difficult to prove—fact is that any closed Jordan curve Γ partitions the plane into two regions: a bounded, open region $\text{int } \Gamma$ and an unbounded, open region $\text{ext } \Gamma$.

Last, let us define the notion of a phase function $\phi: I \rightarrow \mathbb{R}$ which, simply put, describes the phase of the first degree polynomial $s - z$ as s traverses Γ . Specifically, for $z \notin \Gamma$ define $\phi: I \rightarrow \mathbb{R}$ to be any continuous function such that

$$\phi(t) \equiv \arg(\Gamma(t) - z) \pmod{2\pi}. \quad (2)$$

For $z \in \Gamma \setminus \{\Gamma(0)\}$ take this same definition but at $t' \in (0, 1)$ where $\Gamma(t') - z = 0$, only require ϕ to be right continuous and $\lim_{t \rightarrow t'^-} \phi(t) - \phi(t') \in [0, 2\pi)$. This minor complication arises, of course, because (2)'s right-hand side is undefined at $t = t'$. For technical reasons ϕ will be left undefined in the $z = \Gamma(0)$ case.

From Fig. 2 it should be apparent that when $\text{int } \Gamma$ is convex at least one phase function ϕ —and actually, infinitely many phase functions differing by multiples of 2π —exists for any given $z \neq \Gamma(0)$. Further, ϕ will have certain properties that depend only on whether z is inside, on, or outside Γ .

Case 1: $z \in \text{int } \Gamma$. Here ϕ increases continuously and monotonically by 2π (consider $z = z_-$ in Fig. 2). Hence, $V\phi = 2\pi$ and $\phi(1^-) - \phi(0) = 2\pi$.

Case 2: $z \in \Gamma \setminus \{\Gamma(0)\}$. ϕ increases continuously and monotonically by some amount $\theta_0 \leq \pi$ except at the point $t' \in (0, 1)$, $\Gamma(t') - z = 0$, where it makes a downward jump of height θ_0 (consider $z = z_0$ in Fig. 2). Hence, $V\phi = 2\theta_0 \leq 2\pi$ and $\phi(1^-) - \phi(0) = 0$.

Case 3: $z \in \text{ext } \Gamma$. ϕ increases continuously and monotonically by some amount $\theta_+ < \pi$ over a portion of the interval I (consider $z = z_+$ and the counterclockwise $B \rightarrow A$ portion of Γ in Fig. 2) while on the other portion it decreases continuously and monotonically by θ_+ (consider the $A \rightarrow B$ portion of Γ). Hence, $V\phi = 2\theta_+ < 2\pi$ and $\phi(1^-) - \phi(0) = 0$.

Now, from these observations we can prove the "if" part of Theorem 1. Let p be of degree $0 \leq d \leq n$ and have (not necessarily distinct) zeros $\{z_k\}$; so

$$p(s) = \sum_{j=0}^n \alpha_j s^j = \alpha_d \prod_{k=1}^d (s - z_k).$$

Substituting $\Gamma(t)$ for s and letting ϕ_k be an arbitrary phase function associated with z_k , we see

$$\begin{aligned} \arg p(\Gamma(t)) &\equiv \arg \alpha_d + \sum_{k=1}^d \arg(\Gamma(t) - z_k) \\ &\equiv \Phi(t) \pmod{2\pi} \end{aligned}$$

$$\text{where } \Phi(t) = \arg \alpha_d + \sum_{k=1}^d \phi_k(t).$$

Let $0 = t_1 < t_2 < \dots < t_m < 1$ be such that $\Gamma(t_k) = s_k$, $1 \leq k \leq m$, and note (1c) implies $\{\Gamma(t_k)\} \cap \{z_k\} = \emptyset$; so, (1d) implies for all k and some $\zeta_k \in \mathbb{Z}$, $\Phi(t_k) = \theta_k + 2\pi\zeta_k$. Thus, $\Phi(1^-) = \Phi(0) + 2\pi N = \theta_1 + 2\pi(N + \zeta_1)$ where N is the number of zeros z_k lying in $\text{int } \Gamma$. Now, observe

$$2\pi n = 2\pi n + \theta_1 - \theta_m + \sum_{k=1}^{m-1} \theta_{k+1} - \theta_k \quad (3)$$

$$\leq |2\pi n + \theta_1 - \theta_m| + \sum_{k=1}^{m-1} |\theta_{k+1} - \theta_k| \quad (4)$$

$$\begin{aligned} &\leq |2\pi n + \theta_1 - \theta_m + 2\pi(N + \zeta_1 - \zeta_m - n)| \\ &\quad + \sum_{k=1}^{m-1} |\theta_{k+1} - \theta_k + 2\pi(\zeta_{k+1} - \zeta_k)| \end{aligned} \quad (5)$$

$$= |\Phi(1^-) - \Phi(t_m)| + \sum_{k=1}^{m-1} |\Phi(t_{k+1}) - \Phi(t_k)| \quad (6)$$

$$\begin{aligned} &\leq V\Phi \leq \sum_{k=1}^d V\phi_k \\ &\leq 2\pi d \leq 2\pi n \end{aligned} \quad (7)$$

where in going from (4) to (5) we used (1a), (1b) and the fact n , N , and ζ_k are integers. Since the left-hand side of (3) and right-hand side of (7) are equal, we would have a contradiction if the inequality between (4) and (5) were strict. Hence, necessarily $N + \zeta_1 - \zeta_m - n = 0$ and $\zeta_{k+1} - \zeta_k = 0$, $1 \leq k < m$. Summing these equations shows $N = n$. Thus, as we set out to prove, p is of degree n and has all its zeros in Γ .

It now simply remains to prove the "only if" part of Theorem 1. Let p be of degree n and have all of its zeros in $\text{int } \Gamma$. If $n = 0$ trivially $p(s) = \alpha_0 \neq 0$, and Theorem 1 is satisfied for $m = \theta_1 = \arg \alpha_0$, and any $s_1 \in \Gamma$. So, suppose $n > 0$. By Nyquist theorem, the point $p(s)$ will revolve $2\pi n$ radians counterclockwise about the origin as s traverses Γ . Hence, if we choose $m = 2n + 1$ and $\theta_k = \pi(1 - 1/3n)(k - 1)$, $1 \leq k \leq m$, by the continuity of $p(s)$ we will be able to determine $\{s_k\}$ satisfying the theorem. A

interesting point here is that $m = 2n + 1$ is the smallest m for which the theorem can be satisfied. Q.E.D.

With little effort Theorem 1 and its proof can be generalized to certain, nonconvex, simply connected regions as follows. Suppose Γ is an arbitrary closed Jordan curve and $\nu > 0$ a constant such that for each $z \neq \Gamma(0)$ there exists a phase function ϕ with $\forall \phi \leq \nu$ (note, ϕ may not exist if $z \in \Gamma$ and in some neighborhood of z , Γ "oscillates too rapidly"). When $\text{int } \Gamma$ is convex, we saw there always exists a ϕ with $\forall \phi \leq \nu = 2\pi$. When $\text{int } \Gamma$ is not convex, however, we must take $\nu > 2\pi$. Now, to make the above proof go through, the right-hand sides of (1ab) must be replaced by some $\kappa > 0$ where κ, ν , and n satisfy the inequality $\nu n - 2\pi n \leq 2\pi - 2\kappa$. For then, noting that the right-hand side of (7) is νn , we must conclude $N + \zeta_1 - \zeta_m - n = 0$ and $\zeta_{k+1} - \zeta_k = 0$, $1 \leq k < m$, to avoid making the gap between (4) and (5) so large as to yield a contradiction. As before, summing these equations shows $N = n$.

The above inequalities imply $n \leq (2\pi - 2\kappa)/(\nu - 2\pi)$ and $\kappa < \pi$ (when $n > 0$). So, generalizing Theorem 1 in this way requires placing an upper bound on the degree n of p (perhaps requiring $n = 0$) and making the right-hand sides of (1ab) strictly less than π . In our opinion, these conditions are rather ugly, and so, as matter of aesthetics, we chose to state Theorem 1 for only convex regions.

Note, a more useful generalization of Theorem 1 is to unbounded convex regions. This is readily accomplished by approximating a given unbounded region by a suitably large bounded region as was done in [2]. Also, in certain cases where conjugate symmetry is present Theorem 1 can be specialized to require roughly half as many evaluations of p as would otherwise be needed [1], [2].

III. THE FINITE INCLUSIONS THEOREM

A problem of fundamental importance in robust control theory is that of proving a family of linear differential equations has only solutions which decay and/or oscillate within a prescribed range of rates. Via the Laplace transform this problem can be reduced to proving a family of polynomials has all its zeros inside a prescribed region of the complex plane. With this problem in mind we now state the following, very useful corollary to Theorem 1.

Theorem 2—Finite Inclusions Theorem: Let $p(s; q) = \sum_{j=0}^n \alpha_j (q)s^j$, $q \in Q$, where Q is an arbitrary set, $n \geq 0$, and $\alpha_j: Q \rightarrow \mathbb{C}$. Further, let $\Gamma \subset \mathbb{C}$ be a closed Jordan curve such that $\text{int } \Gamma$ is convex (see Fig. 1). Then, for all $q \in Q$, $p(\cdot; q)$ is of degree n and has all its zeros in $\text{int } \Gamma$ if there exist $m \geq 1$ intervals $(a_k, b_k) \subset \mathbb{R}$ and a counterclockwise sequence of points $s_k \in \Gamma$, $1 \leq k \leq m$, such that

$$\forall_{1 \leq k < m} \max\{b_{k+1} - a_k, b_k - a_{k+1}\} \leq \pi \quad (8a)$$

$$\max\{b_m - (a_1 + 2\pi n), (b_1 + 2\pi n) - a_m\} \leq \pi \quad (8b)$$

$$\forall_{1 \leq k \leq m} p(s_k; Q) \subset S_k = \{re^{i\theta} | r > 0, \theta \in (a_k, b_k)\}. \quad (8c)$$

Proof: For each $q \in Q$, (8c) implies for all k , $p(s_k; q) \neq 0$ and there exist $\theta_k \in (a_k, b_k)$ such that $\arg p(s_k; q) \equiv \theta_k \pmod{2\pi}$. Further, from the inequality

$$\forall_{\substack{\theta \in (a,b) \\ \theta' \in (c,d)}} |\theta - \theta'| < \sup_{\substack{\theta \in (a,b) \\ \theta' \in (c,d)}} |\theta - \theta'| = \max\{d - a, b - c\}$$

we see conditions (8a), (8b) imply $\forall_{1 \leq k < m} |\theta_{k+1} - \theta_k| < \pi$ and $2\pi n + \theta_1 - \theta_m < \pi$. Hence, for each $q \in Q$ Theorem 1 implies $p(\cdot; q)$ is of degree n and has all its zeros in $\text{int } \Gamma$. Q.E.D.

For a convex region $\text{int } \Gamma$ Theorem 2 roughly says that if we can find a sequence of open sectors S_k "spaced" no more than π radians apart and revolving a net $2\pi n$ radians about the origin and if we

can find a counterclockwise sequence of points $s_k \in \Gamma$ such that the value sets $p(s_k; Q) = \{p(s_k; q), q \in Q\}$ lie in these sectors, then each member of the family of polynomials $p(\cdot; q)$, $q \in Q$, is of degree n and has all its zeros in $\text{int } \Gamma$.

In general, it should be noted that Theorem 2 is only a sufficient condition for a family of polynomials to have its zeros in a given, convex region. When $p(s; q)$ is continuous and affine in $q \in Q$ and Q is a compact, convex set, however, it is also necessary (i.e., the word "if" in Theorem 2 may be replaced by "if and only if.") To see this, suppose $p(s; q)$ is continuous and affine in q , Q is a compact, convex set, and for all $q \in Q$, $p(\cdot; q)$ is a degree n polynomial with all its zeros in $\text{int } \Gamma$. First note, $p(s; Q)$ is a compact, convex set that revolves n times counterclockwise about the origin without touching the origin as s traverses Γ (since for arbitrary $q \in Q$, $p(s; q)$ revolves in this fashion). By the above facts and the continuous deformation of $p(s; Q)$ as s traverses Γ , we may define two continuous functions $a, b: [0, 1] \rightarrow \mathbb{R}$, such that for all $t \in I$, $\{re^{i\theta} | r > 0, \theta \in [a(t), b(t)]\}$ is the smallest sector containing $p(\Gamma(t); Q)$, and so, $0 \leq b(t) - a(t) < \pi$, $a(1) = a(0) + 2\pi n$, and $b(1) = b(0) + 2\pi n$. Since a continuous function defined on a compact set achieves its supremum and is uniformly continuous on that set, we see that $c = \sup b(t) - a(t) = \max b(t) - a(t) < \pi$ and that a and b are uniformly continuous. Thus, for an arbitrary $0 < \epsilon < (\pi - c)/2$ we can choose a sufficiently fine partition $\{t_k\}$ of I such that conditions (8a), (8b) are satisfied with $a_k = a(t_k) - \epsilon$ and $b_k = b(t_k) + \epsilon$, and by construction, (8c) is satisfied with $s_k = \Gamma(t_k)$.

REFERENCES

- [1] R. D. Kaminsky and T. E. Djaferis, "A novel approach to the analysis and synthesis of controllers for parametrically uncertain systems," *Proc. 1992 IEEE CDC*, Tucson, AZ, to appear, *IEEE Trans. Automat. Contr.*, pp. 333–338.
- [2] —, "The finite inclusions theorem," in *Proc. 1993 IEEE CDC*, San Antonio, TX, pp. 508–518.
- [3] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*, R. A. Silverman, translator and Ed. New York: Dover, 1970.
- [4] A. Rantzer, "Parametric uncertainty and feedback complexity in linear control systems," Ph.D. thesis, Division of Optimization and Systems Theory, Royal Institute of Technology, Sweden, Dec. 1991.

Further Results on Rational Approximations of \mathcal{L}^1 Optimal Controllers

Zi-Qin Wang, Mario Sznajder, and Franco Blanchini

Abstract—The continuous-time persistent disturbance rejection problem (\mathcal{L}^1 optimal control) leads to nonrational compensators, even for SISO systems [4], [7], [8]. As noted in [4], the difficulty of physically implementing these controllers suggests that the most significant applications of the continuous time \mathcal{L}^1 theory is to furnish bounds for the achievable performance of the plant. Recently, two different rational approximations of the optimal \mathcal{L}^1 controller were developed by Ohta *et al.* [6] and by Blanchini and Sznajder [1]. In this paper we explore the connections between these two approximations. The main result of the paper shows that both approximations belong to the same subset Ω_τ of the set of rational approximations, and that the method proposed in [1] gives the best approximation, in the sense of providing the tightest upper bound of the approximation error, among the elements of this subset. Additionally, we exploit the structure of the dual to the \mathcal{L}^1 optimal control problem to obtain rational approximations with approximation error smaller than a prespecified bound.

I. INTRODUCTION

A large number of control problems involve designing a controller capable of stabilizing a given linear time invariant system while minimizing the worst case response to some exogenous disturbances. This problem is relevant for instance for disturbance rejection, tracking and robustness to model uncertainty (see [2] and references therein). When the exogenous disturbances are modeled as bounded energy signals and performance is measured in terms of the energy of the output, this problem leads to the well known \mathcal{H}_∞ theory. The case where the signals involved are persistent bounded signals leads to the \mathcal{L}^1 optimal control theory, formulated and further explored by Vidyasagar [7], [8] and solved by Dahleh and Pearson both in the discrete- [3], [5] and continuous-time [4] cases.

The \mathcal{L}^1 theory is appealing because it directly incorporates time-domain specifications. Moreover, it furnishes a complete solution to the robust performance problem (see [2] for a good tutorial and a list of relevant references). However, in contrast with the discrete time l^1 theory, the solution for the continuous-time \mathcal{L}^1 optimal control problem leads to nonrational compensators, even for SISO systems. As noted in [4], the difficulty of physically implementing these controllers suggests that the most significant application of the continuous time \mathcal{L}^1 theory is to provide performance bounds for the plant. Recently, two rational approximations to the optimal \mathcal{L}^1 controller were developed independently [6], [1]. Although these approximations are based upon different techniques ([6] follows an algebraic approach while [1] exploits the properties of the Euler approximating set), they seem to be strongly connected [1]. Noteworthy, they yield closed-loop plants with the same pole structure.

In this paper we explore the connection between these approaches. The main result of the paper shows that both belong to the same subset Ω_τ of the set of admissible rational approximations, and that the method proposed in [1] gives the best approximation (in the sense of providing the tightest upper bound of the error) among the elements of this set. Additionally, by exploiting the structure of the dual to

the \mathcal{L}^1 optimal control problem we furnish a procedure to compute rational approximations with error smaller than a prespecified bound ϵ , and we show that the approximation error $\rightarrow 0$ as $O(\tau)$.

The paper is organized as follows. In Section II we introduce the notation to be used and we restate the main results concerning the \mathcal{L}^1 problem and its rational approximations. Section III contains the majority of the theoretical results. Here we compare the two approximation methods and we show that, in a sense, the method proposed in [1] yields the best rational approximation. In Section IV we present a simple design example and we compare the optimal \mathcal{L}^1 controller with its rational approximations. Finally, in Section V we summarize our results.

II. PRELIMINARIES

A. Notation and Definitions

R_+ denotes the set of nonnegative real numbers. $\mathcal{L}^\infty(R_+)$ denotes the space of measurable functions $f(t)$ equipped with the norm: $\|f\|_\infty = \text{ess} \cdot \sup_{R_+} |f(t)|$. $\mathcal{L}^1(R_+)$ denotes the space of Lebesgue integrable functions on R_+ equipped with the norm $\|f\|_1 \triangleq \int_0^\infty |f(t)| dt < \infty$. Similarly, l_1 denotes the space of absolutely summable sequences $h = \{h_i\}$ equipped with the norm $\|h\|_1 \triangleq \sum_{i=0}^\infty |h_i| < \infty$. \mathcal{RL}^1 denotes the subspace of \mathcal{L}^1 formed by matrices with real rational Laplace transform. A denotes the space whose elements have the form

$$h = h^L(t) + \sum_{k=0}^\infty h_k^I \delta(t - t_k)$$

where $h^L(t) \in L_1(R_+)$, $\{h_k^I\} \in l_1$ and $t_k \geq 0$, equipped with the norm $\|h\|_A \triangleq \|h^L\|_{L_1} + \|h^I\|_{l_1}$. Given a function $f(t) \in \mathcal{L}^1$ we denote its Laplace transform by $F(s) \in \mathcal{L}_\infty$; similarly, given $h \in A$, we denote its Laplace transform by $H(s)$. By a slight abuse of notation, we denote as $\|F(s)\|_1 \triangleq \|f(t)\|_1$ and $\|H(s)\|_A = \|h\|_A$. Throughout the paper we use packed notation to represent state-space realizations, i.e.,

$$G(s) = C(sI - A)^{-1}B + D \triangleq \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right).$$

Definition 1: Consider the continuous time system $G(s)$. Its Euler approximating system (EAS) is defined as the following discrete time system

$$G^E(z, \tau) = \left(\begin{array}{c|c} I + \tau A & \tau B \\ \hline C & D \end{array} \right). \quad (1)$$

From this definition it is easily seen that we can obtain the EAS of $G(s)$ by the simple variable transformation $s = (z - 1)/\tau$, i.e.,

$$G^E(z, \tau) = G\left(\frac{z-1}{\tau}\right).$$

On the other hand, for any given τ we can relate a discrete time system to a continuous system by the inverse transformation $z = 1 + \tau s$. It is obvious that the discrete time system is, in fact, the EAS of the continuous time system obtained in this form.

Definition 2: Consider a system of the form

$$\Phi(s) = T_1(s) + T_2(s)Q(s)$$

where $T_2(s)$ has all its zeros $\{z_1, z_2, \dots, z_n\}$ in the open right-half plane and where, for simplicity, we assume that all the zeros are

Manuscript received October 8, 1993; revised July 12, 1994. This work was supported in part by the NSF under Grant ECS-9211169.

Z.-Q. Wang and M. Sznajder are with the Department of Electrical Engineering, The Pennsylvania State University, University Park, PA 16802 USA.

F. Blanchini is with the Dipartimento di Matematica e Informatica, Università degli Studi di Udine, 33100, Udine, Italy.

IEEE Log Number 9407003.

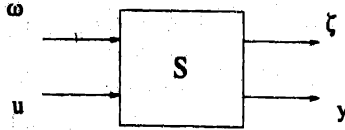


Fig. 1. The generalized plant.

distinct. $\Phi(s)$ is said to be admissible if it is stable and satisfies the interpolation conditions

$$\Phi(z_k) = T_1(z_k), \quad k = 1, \dots, n.$$

B. The \mathcal{L}^1 Optimal Control Problem

Consider the system shown in Fig. 1, where S represents the system to be controlled; the scalar signals $\omega \in \mathcal{L}^\infty$ and u represent an exogenous disturbance and the control action respectively; and where ζ and y represent the output subject to performance constraints and the measurements available to the controller, respectively. As usual we will assume, without loss of generality, that any weights have been absorbed in the plant S . Then, the \mathcal{L}^1 optimal control problem can be stated as: Given the system (S) find an internally stabilizing controller $u(s) = K(s)y(s)$ such that the worst case (over the set of all $\omega(t) \in \mathcal{L}^\infty$, $\|\omega\|_\infty \leq 1$) maximum amplitude of the performance output $\zeta(t)$ is minimized.

By using the YJBK parameterization of all stabilizing controllers [4], [9], the problem can be cast into the following model matching form

$$\mu_0 = \inf_{K \text{ stabilizing}} \|\Phi(s)\|_A = \inf_{Q \in A} \|T_1(s) + T_2(s)Q(s)\|_A \quad (2)$$

where T_1, T_2 are rational stable transfer functions.

Next, we recall the main result of [4], showing that a solution to the \mathcal{L}^1 optimal control problem can be found by solving a semi-infinite linear programming problem.

Theorem 1 (Dahleh and Pearson [4]): Let $T_2(s)$ have n zeros z_i in the open right-half plane and no zeros on the $j\omega$ axis. Then

$$\begin{aligned} \mu_0 &= \inf_{Q \in A} \|T_1(s) + T_2(s)Q(s)\|_A \\ &= \max_{\alpha_j} \left[\sum_{i=1}^n \alpha_i \operatorname{Re}\{T_1(z_i)\} + \sum_{i=1}^n \alpha_{i+n} \operatorname{Im}\{T_1(z_i)\} \right] \end{aligned} \quad (3)$$

subject to

$$\left| r(t) \right| = \left| \sum_{i=1}^n \alpha_i \operatorname{Re}\{e^{-z_i t}\} + \sum_{i=1}^n \alpha_{i+n} \operatorname{Im}\{e^{-z_i t}\} \right| \leq 1, \quad \forall t \in R_+. \quad (4)$$

Furthermore, the following facts hold: i) the extremal functional $r^*(t)$ equals 1 at only finite points: t_1, \dots, t_m ; ii) an optimal solution $\Phi(s) = T_1(s) + T_2(s)Q(s)$ to the left side problem always exists; and iii) the optimal ϕ has the following form

$$\phi = \sum_{i=1}^m \phi_i \delta(t - t_i), \quad t_i \in R_+, \quad m \text{ finite} \quad (5)$$

and satisfies the following conditions

- a) $\phi_i r^*(t_i) \geq 0$;
- b) $\sum_{i=1}^m |\phi_i| = \mu_0$;
- c) $\sum_{i=1}^m \phi_i e^{-z_k t_i} = T_1(z_k), \quad k = 1, \dots, n.$

Remark 1: It was shown in [4] that we need to satisfy constraints (4) only for all $t \leq t_{\max}$ where t_{\max} is finite and can be determined a priori. Even so, there are still infinite constraints, and therefore the dual problem is a semi-infinite linear programming problem.

C. Rational Approximations to the Optimal \mathcal{L}^1 Controller

From (5) it follows that, unlike in the discrete-time case, the \mathcal{L}^1 optimal controller is irrational even if the plant is rational. Prompted by the difficulty in physically implementing a controller with a nonrational transfer function, two rational approximation methods have been recently developed by Ohta *et al.* [6] and by Blanchini and Sznajder [1]. In the sequel we briefly review these results. For brevity, we refer to the former as the OMK method and to the latter as the EAS method.

Theorem 2 (Ohta *et al.*, [6]): Let

$$T_1(s) = \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} \quad \text{and} \quad T_2(s) = \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix}$$

be minimal realizations. Define

$$\begin{aligned} L &= B_2 D_2^{-1} \\ \hat{A} &= A_2 - L C_2 \\ M &= L D_1 + R_1 B_1 \end{aligned}$$

where R_1 is the unique solution of the linear matrix equation

$$\hat{A} R_1 - R_1 A_1 = L C_1.$$

Then, there exist finite sets $\{t_1, t_2, \dots, t_m\}$ and $\{\phi_1, \phi_2, \dots, \phi_m\}$ such that

$$M = \sum_{i=1}^m \phi_i \exp(-\hat{A} t_i) L \quad (6)$$

and $\mu_0 = \sum_{i=1}^m |\phi_i|$. For $\tau > 0$, define $N(t_i, \tau) \triangleq$ the smallest integer larger than or equal to t_i/τ , $i = 1, 2, \dots, m$, and $N(\tau) \triangleq N(t_m, \tau)$. Finally, denote by $\phi(\tau)$ the minimizer of $\|\phi(\tau) - \phi\|_2$ subject to

$$M = \sum_{i=1}^m \phi_i(\tau) (I + \tau \hat{A})^{-N(t_i, \tau)} L \quad (7)$$

where $\phi(\tau) = [\phi_1(\tau), \dots, \phi_m(\tau)]$, and $\phi = [\phi_1, \dots, \phi_m]$. Consider the rational system $\Phi(s, \tau)$ with the following state-space realization

$$\Phi(s, \tau) = \begin{pmatrix} A(\tau) & B_4(\tau) \\ C_3(\tau) & D_{34}(\tau) \end{pmatrix}$$

where the matrices A, B_4, C_3, D_{34} are defined as follows

$$A(\tau) = \tau^{-1} \begin{pmatrix} -1 & 1 & 0 & \dots & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ 0 & 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & -1 & 1 \\ 0 & \dots & \dots & 0 & 0 & -1 \end{pmatrix}, \quad N(\tau) \text{ by } N(\tau)$$

$B_4(\tau)$: its k th element $\begin{cases} 0, & \text{if } k \notin N(t_i, \tau), i = 1, 2, \dots, m \\ \phi_i(\tau), & \text{if } k = N(t_i, \tau) \end{cases}$

$$C_3(\tau) = (\tau^{-1} \quad 0 \quad \dots \quad 0), \quad 1 \text{ by } N(\tau)$$

$$D_{34}(\tau) = \begin{cases} 0, & \text{if } N(t_1, \tau) \neq 0 \\ \phi_1(\tau), & \text{if } N(t_1, \tau) = 0. \end{cases}$$

Then, as $\tau \rightarrow 0$, we have that $\Phi(s, \tau) \rightarrow \Phi_{\text{OPT}}(s)$ uniformly in the wide sense in the open half plane $\operatorname{Re}(s) > -\sigma$ for some $\sigma > 0$; and $\|\Phi\|_A$, as well as its upper bound $\gamma = \sum_{i=1}^m |\phi_i(\tau)|$, converge to μ_0 .

Remark 2: As we will show later, (6) and (7) are just another version of the interpolation condition.

Next, we recall the main result of [1] showing that the \mathcal{L}^1 norm of a stable transfer function is bounded above by the l^1 norm of its EAS. Moreover, this bound can be made arbitrarily tight by taking the parameter τ in (1) small enough. This result is the basis for the approximation procedure proposed in [1].

Theorem 3 (Blanchini and Sznajder [1]): Consider a continuous time system with rational Laplace transform $\Phi(s)$ and its EAS, $\Phi^E(z, \tau)$. If $\Phi^E(z, \tau)$ is asymptotically stable, then $\Phi(s)$ is also asymptotically stable and such that

$$\|\Phi(s)\|_1 \leq \|\Phi^E(z, \tau)\|_1.$$

Conversely, if $\Phi(s)$ is asymptotically stable and such that $\|\Phi(s)\|_1 \triangleq \mu_c$, then for all $\mu > \mu_c$ there exists $\tau^* > 0$ such that for all $0 < \tau \leq \tau^*$, $\Phi^E(z, \tau)$ is asymptotically stable and such that $\|\Phi^E(z, \tau)\|_1 \leq \mu$.

Theorem 4 [1]: Consider a strictly decreasing sequence $\tau_i \rightarrow 0$, and define

$$\mu_i \triangleq \inf_{\text{stabilizing } K} \|\Phi_{cl}^E(z, \tau_i)\|_1$$

where $\Phi_{cl}^E(z, \tau_i)$ denotes the closed-loop transfer function. Then the sequence μ_i is nonincreasing and such that $\mu_i \rightarrow \mu_0$, the optimal \mathcal{L}^1 cost.

Corollary: A suboptimal rational solution to the \mathcal{L}^1 optimal control problem for continuous time systems, with cost arbitrarily close to the optimal cost, can be obtained by solving a discrete-time l_1 optimal control problem for the corresponding EAS. Moreover, if $K(z)$ denotes the optimal l_1 compensator for the EAS, the suboptimal \mathcal{L}^1 compensator is given by $K(\tau s + 1)$.

III. ANALYSIS OF THE DIFFERENT RATIONAL APPROXIMATIONS

In this section we analyze the rational approximations generated by the OMK and EAS methods. The main result shows that both approximations belong to a certain subset Ω_τ of the set of rational approximations, and that the EAS method generates the best approximation among the elements of this subset. We begin by showing that the two expressions for matrix M in Theorem 2 are just another version of the interpolation conditions.

A. Characterization of All Rational Approximations

Lemma 1: For a closed-loop system of the form

$$\Phi(s) = \sum_{i=1}^q \phi_i e^{-t_i s}, \quad t_i \in \mathcal{R}_+$$

the following two conditions are equivalent

- $\Phi(z_k) = \sum_{i=1}^q \phi_i e^{-z_k t_i} = T_1(z_k), \quad k = 1, \dots, n.$
- $M = \sum_{i=1}^q \phi_i \exp(-\hat{A} t_i) L.$

Moreover, we have $\|\Phi(s)\|_A = \gamma \triangleq \sum_{i=1}^q |\phi_i|$.

Proof: b) \Rightarrow a) can be proved following the proof procedure of [6, lemma 2] by simply replacing Φ_{OPT} with Φ . Similarly, the fact that b) is necessary for a) to hold can also be concluded from the proof. The expression for $\|\Phi(s)\|_A$ follows from direct calculations.

In the next lemma we give a complete characterization of all rational approximations.

Lemma 2: For any rational closed-loop system

$$\Phi(s) = \sum_{j=1}^{q_j} \sum_{i=1}^{q_i} \phi_{ij} (1 + \lambda_j s)^{-N_{ij}}$$

where $\text{Re}(\lambda_j) > 0$ and N_{ij} integers, the following two conditions are equivalent

- $\Phi(z_k) = \sum_{j=1}^{q_j} \sum_{i=1}^{q_i} \phi_{ij} (1 + \lambda_j z_k)^{-N_{ij}} = T_1(z_k), \quad k = 1, \dots, n.$
- $M = \sum_{j=1}^{q_j} \sum_{i=1}^{q_i} \phi_{ij} (I + \lambda_j \hat{A})^{-N_{ij}} L.$

Moreover, we have $\|\Phi(s)\|_A \leq \gamma \triangleq \sum_{j=1}^{q_j} \sum_{i=1}^{q_i} |\phi_{ij}|$.

Proof: a) \Leftrightarrow b) can be proved using the same idea. The calculations, through straightforward, are tedious, and are omitted here for space reasons. By direct calculation we have

$$\begin{aligned} \|\Phi(s)\|_A &= \int_0^\infty \left| \sum_{j=1}^{q_j} \sum_{i=1}^{q_i} \phi_{ij} \frac{1}{(N_{ij}-1)! \lambda_j^{N_{ij}}} t^{N_{ij}-1} e^{-t/\lambda_j} \right| dt \\ &\leq \sum_{j=1}^{q_j} \sum_{i=1}^{q_i} |\phi_{ij}| \int_0^\infty \frac{1}{(N_{ij}-1)! \lambda_j^{N_{ij}}} t^{N_{ij}-1} e^{-t/\lambda_j} dt \\ &= \gamma \triangleq \sum_{j=1}^{q_j} \sum_{i=1}^{q_i} |\phi_{ij}|. \quad \square \end{aligned}$$

B. Comparison of the OMK and the EAS Rational Approximations

Lemma 2 gives a characterization of all rational admissible closed-loop systems. All these closed-loop systems can be thought as candidate rational approximations of the \mathcal{L}^1 -optimal closed-loop system. In the sequel we concentrate on a specific subset Ω_τ and we show that both the OMK and the EAS methods generate approximations that belong to this subset. For $\tau > 0$, define

$$\begin{aligned} \Omega_\tau &= \left\{ \Phi(s) = \sum_{i=1}^q \phi_i(\tau) (1 + \tau s)^{-N_i} : M \right. \\ &= \left. \sum_{i=1}^q \phi_i(\tau) (I + \tau \hat{A})^{-N_i} L \right\}. \end{aligned} \quad (8)$$

By direct calculation, the closed-loop system obtained by OMK methods is

$$\Phi(s) = \sum_{i=1}^m \phi_i(\tau) (1 + \tau s)^{-N_i(\tau)}.$$

Suppose that the l_1 -optimal closed-loop system for the EAS is given by

$$\Phi^E(z) = \sum_{i=1}^q \phi_i^E(\tau) z^{-N_i^E}$$

then the closed-loop system obtained using the EAS method is

$$\Phi(s) = \sum_{i=1}^q \phi_i^E(\tau) (1 + \tau s)^{-N_i^E}.$$

It follows that the approximations generated by both methods belong to the set Ω_τ , with a specific $\{N_i\}$ determined by each method.

Remark 3: In the OMK method, $\{N_i(t_i, \tau)\}$ depend directly on $\{t_i\}$, and hence on the \mathcal{L}^1 optimal closed-loop system. Hence obtaining a rational approximation requires solving the \mathcal{L}^1 optimal control problem first. However, as pointed out in Remark 1, solving exactly this problem entails solving a semi-infinite linear programming problem. The EAS method requires only solving a discrete l_1 optimal control problem, which is considerably easier, since only finite-dimensional linear programming is involved.

Remark 4 Note that additional OMK-like rational approximations can be chosen among the elements of Ω_τ by simply modifying the rule for selecting $N(t, \tau)$. For instance, N could be selected as the largest integer smaller than or equal to t_i/τ or $t/\tau + 0.5$. Clearly, the convergence property also holds for these approximations.

As it is shown in Theorem 6, the EAS method can be interpreted as approximating the original optimization problem as opposed to directly approximating its irrational solution. This makes it quite unique. Besides the computational advantages, we show in the sequel that the EAS method has two other important merits.

Theorem 5 The rational approximation of the \mathcal{L}^1 optimal controller given by the EAS method is the best one in the set Ω_τ in the sense that it leads to the smallest upper bound γ .

To prove Theorem 5, we need to prove first the following results.

Lemma 3 Consider the following discrete time systems

$$T^I = \left(\begin{array}{c|c} A_{1I} & B_{1I} \\ \hline C_{1I} & D_{1I} \end{array} \right) \triangleq \left(\begin{array}{c|c} I + \tau A_1 & \tau B_1 \\ \hline C_1 & D_1 \end{array} \right)$$

and

$$T^I = \left(\begin{array}{c|c} A_{2I} & B_{2I} \\ \hline C_{2I} & D_{2I} \end{array} \right) \triangleq \left(\begin{array}{c|c} I + \tau A_2 & \tau B_2 \\ \hline C_2 & D_2 \end{array} \right)$$

Define

$$\begin{aligned} I_I &= B_{2I} D_{1I}^{-1} \\ A_I &= A_{2I} - I_I C_{1I} \\ M_I &= I_I D_{1I} + A_I^{-1} R_{1I} B_{1I} \end{aligned}$$

where R_{1I} is the unique solution of the linear matrix equation

$$A_I^{-1} R_{1I} A_I - R_{1I} + I_I C_{1I} = 0$$

Then we have

$$M_I = \tau M$$

where M is defined in Theorem 2.

Proof First we show that A_I is always invertible. Note that

$$(I^I)^{-1} = \left(\begin{array}{c|c} A_I & -I_I \\ \hline D_{1I}^{-1} C_{1I} & D_{1I}^{-1} \end{array} \right)$$

Since it is assumed that $F(s)$ has only unstable zeros, so does $T^I(z)$. This means all the zeros (poles) of $T^I(z)$ ($(T^I)^{-1}$) have magnitudes larger than 1. Invertibility of A_I follows immediately. Recall now that R_1 is the unique solution of the following linear matrix equation

$$A R_1 - R_1 A - I C_1 = 0$$

We can verify that $A_I^{-1} R_{1I}$ also satisfies the above equation

$$\begin{aligned} A_I A_I^{-1} R_{1I} - A_I^{-1} R_{1I} A_I - I_I C_{1I} \\ = -I_I (R_{1I} - A_I^{-1} R_{1I} A_I - I_I C_{1I}) = 0 \end{aligned}$$

Hence $R_I = A_I^{-1} R_{1I}$ and $M_I = \tau M$.

Lemma 4 Consider the discrete time l_1 optimal control problem for the EAS system

$$\mu_I = \inf_{Q \in \mathcal{H}_1} \|T_1^I + T_2^I Q\|_1$$

A closed-loop system $\Phi^I(z) = \sum_{k=1}^n \phi_k z^{-k}$ is admissible i.e. satisfies the interpolation conditions

$$\Phi^I(z_k) = T_1^I(z_k) \quad k = 1, \dots, n \quad (9)$$

where z_k are the zeros of I_2^I and only if

$$M_I = \sum_{k=1}^n \phi_k A_I^{-k} L_I \quad (10)$$

Proof From the definition of EAS (9) is equivalent to

$$\Phi(z_k) = \sum_{i=1}^l \phi_i (1 + \tau z_k)^{-i} = T_1^I(z_k), \quad k = 1, \dots, n.$$

From Lemma 3 (10) is equivalent to

$$\begin{aligned} M &= \sum_{i=1}^l \phi_i (I + \tau A)^{-i} I \Leftrightarrow \\ \tau M &= \sum_{i=1}^l \phi_i (I + \tau A_2 - \tau I C)^{-i} \tau L \\ M_I &= \sum_{i=1}^l \phi_i A_I^{-i} I \end{aligned} \quad (11)$$

Equivalence of (9) and (10) follows now from Lemma 2.

Note that Lemma 4 is true for any discrete time systems since a discrete time systems can always be thought as an EAS of some continuous time systems.

Proof of Theorem 6 Consider a set of admissible closed-loop systems

$$\Omega_I = \left\{ \Phi^I(z) = \sum_{i=1}^l \phi_i z^{-i} \mid M_I = \sum_{i=1}^l \phi_i A_I^{-i} L_I \right\} \quad (12)$$

From Lemma 3 it follows that conditions (8) and (12) are identical. Therefore for every $\Phi^I(z)$ in Ω_I there exists a corresponding $\Phi(z) = \Phi^I(1 + \tau z)$ in Ω_τ and vice versa. Since

$$\sum_{k=1}^l |\phi_k| = \|\Phi^I\|_1$$

it follows that the closed loop system Φ^I obtained by solving the optimal l_1 control problem for the LAS yields the smallest γ among the elements of the set Ω_I . Hence the rational closed-loop system obtained by EAS methods also has the smallest upper bound γ among the set Ω .

C. The EAS Method Revisited

Although the results of [6] and [1] show that the optimal \mathcal{L}^1 controllers can be approximated arbitrarily close with a rational controller, these results did not provide a way of obtaining an approximation with error smaller than a prescribed bound, rather, they required solving a sequence of problems and checking the approximation error until the desired precision was achieved. In this section we indicate how to select the parameter τ for the EAS method in such a way that the error of the resulting approximation is smaller than a prescribed bound. Moreover, we show that this approximation error converges to 0 as fast as τ .

Theorem 6 Given any $\epsilon > 0$, we can find a τ a priori for the EAS method such that

$$\mu_I \leq \mu_0(1 + \epsilon)$$

Moreover, the approximation error converges to zero as $O(\tau)$.

Proof Consider the optimal l_1 control problem for the EAS

$$\mu_I = \inf_{Q \in \mathcal{H}_1} \|T_1^I(z) + T_2^I(z)Q\|_1 \quad (13)$$

and its dual

$$\mu_I = \max_v \left[\sum_{k=1}^n \alpha_k \operatorname{Re}\{T_1^I(z_k^{-1})\} + \sum_{k=1}^n \alpha_{k+n} \operatorname{Im}\{T_1^I(z_k^{-1})\} \right] \quad (14)$$

subject to

$$\begin{aligned} |T^I(z_k^{-1})| &\triangleq \left| \sum_{i=1}^l \alpha_i \operatorname{Re}\{(z_k^{-1})^{-i}\} + \sum_{i=1}^n \alpha_{i+n} \operatorname{Im}\{(z_k^{-1})^{-i}\} \right| \\ &\leq 1, \quad 0 \leq i \leq 2 \end{aligned} \quad (15)$$

where T_1^E and T_2^E are the EAS of T_1 and T_2 , respectively, and where z_k^E denotes the zeros of T_2^E . From the relationship between the EAS and its corresponding continuous system, the above dual problem is equivalent to

$$\mu_E = \max_{\alpha_j} \left[\sum_{i=1}^n \alpha_i \operatorname{Re}\{T_1(z_i)\} + \sum_{i=1}^n \alpha_{i+n} \operatorname{Im}\{T_1(z_i)\} \right] \quad (16)$$

subject to

$$\left| \sum_{i=1}^n \alpha_i \operatorname{Re}\{(1 + \tau z_i)^{-k}\} + \sum_{i=1}^n \alpha_{i+n} \operatorname{Im}\{(1 + \tau z_i)^{-k}\} \right| \leq 1, \quad k = 0, 1, 2, \dots \quad (17)$$

which can further be thought as an approximation of the dual problem of \mathcal{L}^1 -optimal control problem

$$\mu_0 = \max_{\alpha_j} \left[\sum_{i=1}^n \alpha_i \operatorname{Re}\{T_1(z_i)\} + \sum_{i=1}^n \alpha_{i+n} \operatorname{Im}\{T_1(z_i)\} \right] \quad (18)$$

subject to

$$|\tau(t, \alpha)| \triangleq \left| \sum_{i=1}^n \alpha_i \operatorname{Re}\{e^{-z_i t}\} + \sum_{i=1}^n \alpha_{i+n} \operatorname{Im}\{e^{-z_i t}\} \right| \leq 1, \quad \forall t \in R_+ \quad (19)$$

in the sense that the constraints (19) are firstly sampled at the time interval $t_k = k\tau$ and then the irrational terms $e^{-z_i t_k}$ are approximated by rational terms $(1 + \tau z_i)^{-k}$.

For simplicity, in the sequel we will assume that all the zeros z_i are real as in [4], although the proofs can be easily extended to encompass complex zeros as well.

An upper bound on $\|\alpha\|_1$ for all α satisfying the constraints (15) can be derived as follows. Define the following sets

$$\begin{aligned} S_c &= \left\{ \alpha: \left| \sum_{i=1}^n \alpha_i e^{-z_i t} \right| \leq 1 \forall t \geq 0 \right\} \\ S(\tau) &= \left\{ \alpha: \left| \sum_{i=1}^n \alpha_i (1 + \tau z_i)^{-k} \right| \leq 1 \forall k \geq 0 \right\} \\ R(\tau) &= \left\{ \alpha: \left| \sum_{i=1}^n \alpha_i (1 + \tau z_i)^{-k} \right| \leq 1 \quad k = 0, 1, \dots, n-1 \right\}. \end{aligned} \quad (20)$$

From [1] it can be shown that, if $\tau \leq \bar{\tau}$, then $S_c \subseteq S(\tau) \subseteq S(\bar{\tau})$ and $S(\tau) \subseteq R(\tau)$. Hence

$$\begin{aligned} \sup_{\alpha \in S_c} \|\alpha\|_1 &\leq \sup_{\alpha \in S(\tau)} \|\alpha\|_1 \leq \sup_{\alpha \in S(\bar{\tau})} \|\alpha\|_1 \\ &\leq \sup_{\alpha \in R(\bar{\tau})} \|\alpha\|_1 \leq \|F^{-1}(\bar{\tau})\|_{\infty, 1} \end{aligned} \quad (21)$$

where

$$F(\tau) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ (1 + \tau z_1)^{-1} & (1 + \tau z_2)^{-1} & \dots & (1 + \tau z_n)^{-1} \\ \vdots & \vdots & \dots & \vdots \\ (1 + \tau z_1)^{-n+1} & (1 + \tau z_2)^{-n+1} & \dots & (1 + \tau z_n)^{-n+1} \end{pmatrix}.$$

$\|F^{-1}\|_{\infty, 1}$ denotes the induced norm of F^{-1} from l_n^∞ to l_n^1 , and where $\bar{\tau}$ is fixed.

Note also that constraints (19) will be automatically satisfied for all $t > t_{\max}$ where t_{\max} is finite and can be determined a priori [4]. Given any $t \leq t_{\max}$, assume that the constraints (17) are satisfied and consider any $t \leq t_{\max}$. Selecting k such that $t_k \leq t < t_k + \tau$ we have that

$$\begin{aligned} &\left| \sum_{i=1}^n \alpha_i (1 + \tau z_i)^{-1} - \sum_{i=1}^n \alpha_i e^{-z_i t} \right| \\ &\leq \|\alpha\|_1 \max_i \{ |(1 + \tau z_i)^{-1} - e^{-z_i t}| \} \\ &\leq \|\alpha\|_1 \left(\max_i \{ |(1 + \tau z_i)^{-1} - e^{-z_i t_k}| \} \right. \\ &\quad \left. + \max_i \{ |e^{-z_i t_k} - e^{-z_i t}| \} \right) \\ &\leq \|\alpha\|_1 \left(\max_i \{ t_{\max} \tau^{-1} |(1 + \tau z_i)^{-1} - e^{-z_i \tau}| \} \right. \\ &\quad \left. + \max_i \{ 1 - e^{-z_i \tau} \} \right) \triangleq \epsilon(\tau). \end{aligned} \quad (22)$$

The first inequality is immediate. The second one follows from the triangle inequality. The last one can be proved as follows: If $|a| \leq 1$ and $|b| \leq 1$, then

$$|a^k - b^k| = |a - b| |a^{k-1} + a^{k-2}b + \dots + b^{k-1}| \leq k|a - b|$$

so we have

$$\begin{aligned} |(1 + \tau z_i)^{-1} - e^{-z_i t_k}| &\leq k|(1 + \tau z_i)^{-1} - e^{-z_i \tau}| \\ &\leq t_{\max} \tau^{-1} |(1 + \tau z_i)^{-1} - e^{-z_i \tau}|. \end{aligned}$$

Note that both terms in the parenthesis can be made as small as one desires. So given any $\epsilon > 0$, we can choose a τ such that the right-hand side of the inequality is less than or equal to ϵ . For this value of τ we have

$$|r(t, \alpha)| = \left| \sum_{i=1}^n \alpha_i e^{-z_i t} \right| \leq 1 + \epsilon, \quad \forall t \in R_+.$$

In particular the above inequality holds for α^* which solves the dual problem of EAS. Since

$$\langle T_1, r(t, \alpha^*) \rangle = \sum_{i=1}^n \alpha_i^* T_1(z_i) = \mu_E$$

we have that

$$\mu_0 = \max_{r \neq 0} \frac{\langle T_1, r \rangle}{\|r\|_\infty} = \max_{\|r\|_\infty \leq 1} \frac{\langle T_1, r \rangle}{1 + \epsilon} \geq \frac{\mu_E}{1 + \epsilon}.$$

Finally, the fact that $\epsilon(\tau) = O(\tau)$ follows from considering the Taylor expansion of equation (22).

IV. AN EXAMPLE

Consider the example introduced in [4] and further studied in [1] and [6]. The plant is

$$P(s) = \frac{s-1}{s-2}.$$

The control objective is to minimize $\|\Phi\|_1 = \|PC(1 + PC)^{-1}\|_1$. The optimal closed-loop system is [4]

$$\Phi_{\text{OPT}}(s) = 1.7071 - 4.1213e^{-0.8814s}$$

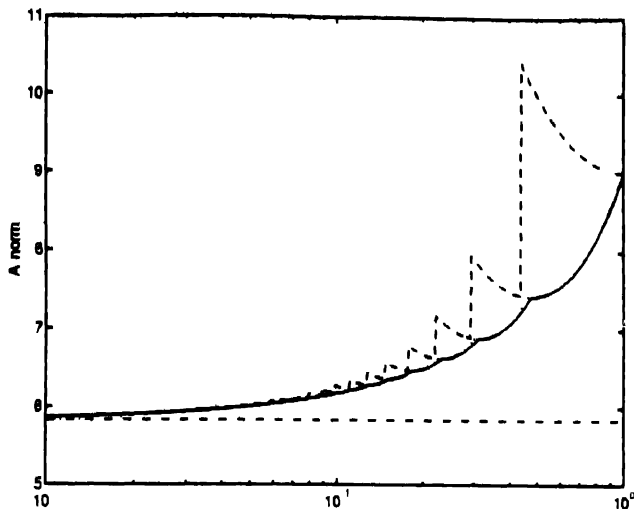


Fig. 2 Upper bound γ versus τ : EAS method—solid line; OMK method—dotted line; and OMK-like method—dashed line.

with an optimal cost $\mu_0 = 5.8284$. For $\tau = 0.45$ the rational closed-loop system obtained using the EAS method is

$$\Phi(s) = 1.8001 - 5.1878(1 + 0.45s)^{-1}$$

with $\gamma = \|\Phi(s)\|_1 = 7.2679$. The OMK method yields

$$\Phi(s) = 2.3947 - 5.0348(1 + 0.45s)^{-1}$$

with $\gamma = \|\Phi(s)\|_1 = 7.4295$. Finally, if we consider the OMK-like approximation obtained by selecting $N(t_i, \tau) \triangleq$ largest integer smaller than or equal to t_i/τ , $i = 1, \dots, m$, we obtain

$$\Phi(s) = 4.222 - 6.122(1 + 0.45s)^{-1}$$

with $\gamma = \|\Phi(s)\|_1 = 10.341$. Fig. 2 shows the upper bounds corresponding to different approximation methods versus τ . For this example the $\| \cdot \|_1$ norm of the closed-loop system coincides with its upper bound in all cases (since there are only 2 interpolation constraints). It is interesting to note that while the bound obtained using the EAS method decreases monotonously with τ (theoretically proved in Theorem 4), those corresponding to the OMK and OMK-like methods do not. An estimated error bound $\epsilon(\tau)$ curves for this example is shown in Fig. 3. It is very close to a straight line if a linear scale is used for the τ axis.

V. CONCLUSIONS

A recent research effort [3]–[5], [7], [8], has led to techniques for designing optimal compensators that minimize the worst case output amplitude with respect to all inputs of bounded amplitude. In the discrete-time SISO case, minimizing the l^1 norm of the closed-loop impulse response yields a rational compensator. Unfortunately, the solution to the continuous-time version of the problem is nonrational. Prompted by the difficulty of physically implementing a system with a nonrational transfer function, rational approximations were recently developed [6], [1].

In this paper we compare these approximations and we show that they are strongly connected. Indeed, both approximations can be considered as elements of the same subset Ω_r of the set of rational approximations.

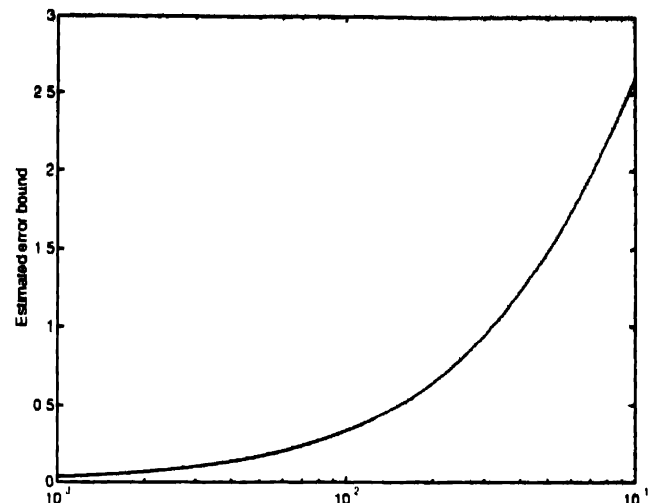


Fig. 3 An estimated error bound.

In Section III-B we show that the EAS method proposed in [1] yields the best approximation (in the sense of providing the tightest upper bound of the error) among the elements of this set.

Finally, in Section III-C we exploit the structure of the dual problem to provide a procedure that allows for selecting the parameter τ for the EAS method to guarantee that the approximation error is smaller than a prespecified bound ϵ . Moreover, we also show that this approximation error $\rightarrow 0$ as $O(\tau)$.

We believe that these results, combined with the features of the EAS method mentioned in [1], namely, the facts that i) it removes the ill-posedness due to the presence of zeros on the imaginary axis; ii) it leads to computationally simple problems; iii) it furnishes a monotonically nonincreasing bound; and iv) it is easily extendable to the MIMO case, make this method an attractive tool for the design of controllers for continuous time systems.

REFERENCES

- [1] F. Blanchini and M. Sznajer, "Rational \mathcal{L}^1 suboptimal compensators for continuous time systems," *IEEE Trans. Automat. Contr.*, vol. 39, no. 7, pp. 1487–1492, 1994.
- [2] M. A. Dahleh and M. H. Khammash, "Controller design for plants with structured uncertainty," *Automatica*, vol. 29, no. 1, pp. 37–56, Jan. 1993.
- [3] M. A. Dahleh and J. B. Pearson, " l^1 optimal feedback controllers for MIMO discrete time systems," *IEEE Trans. Automat. Contr.*, vol. 32, pp. 314–322, Apr. 1987.
- [4] M. A. Dahleh and J. B. Pearson, " \mathcal{L}^1 optimal compensators for continuous time systems," *IEEE Trans. Automat. Contr.*, vol. 32, pp. 889–895, Oct. 1987.
- [5] —, "Optimal rejection of persistent disturbances, robust stability, and mixed sensitivity minimization," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 722–731, Aug. 1988.
- [6] Y. Ohta, H. Maeda, and S. Kodama, "Rational approximation of \mathcal{L}^1 optimal controllers for SISO systems," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1683–1691, Nov. 1992.
- [7] M. Vidyasagar, "Optimal rejection of persistent bounded disturbances," *IEEE Trans. Automat. Contr.*, vol. 31, pp. 527–535, June 1986.
- [8] —, "Further results on the optimal rejection of persistent bounded disturbances," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 642–652, June 1991.
- [9] D. C. Youla, H. A. Jabr, and J. J. Bongiorno, "Modern Wiener-Hopf design of optimal controllers—Part 2: The multivariable case," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 319–338, June 1976.

Supervisory Control of Timed Discrete-Event Systems under Partial Observation

F. Lin and W. M. Wonham

Abstract—This paper extends the authors' previous work on observability of discrete-event systems by taking time into consideration. In a timed discrete-event system, events must occur within their respective lower and upper time bounds. A supervisor can disable, enable, or force some events to achieve a given control objective. We assume that the supervisor does not observe all events, which is often the case in practice. We generalize the concept of observability to timed discrete-event systems and show that it characterizes the existence condition for a supervisor. We also generalize normality, a stronger version of observability, to timed discrete-event systems, which has nice properties that are absent in observability. We then derive conditions under which observability and normality are equivalent. We propose two methods to synthesize a supervisor, a direct approach and an indirect approach. An example is given to illustrate the results.

I. INTRODUCTION

Since the theory of supervisory control was introduced about 10 years ago in [13], many ideas have been developed, notably the concepts of controllability [13], [14], observability [3], [6], [8], and normality [8], [9], [10]. These concepts were, however, defined only for untimed discrete-event systems, in which the occurrence times of events were not modeled. As the theory advances and applications increase, there is a need to generalize these concepts to timed discrete-event systems, in which lower and upper time bounds of events are specified and events must occur within their respective lower and upper time bounds. In [1], controllability is generalized to timed discrete-event systems and supervisory control is discussed under the assumption of full event observation. In practice, however, not all events are observable due to limitations on detection and communication. In such cases, observability becomes relevant, because a supervisor must control the system based on partial observation of events. This paper is devoted to supervisory control under partial observation.

We begin with a brief review of the definitions on timed discrete-event systems and the definition of controllability of [1] in Section II. In Section III, we formalize supervisory control under partial observation and discuss the constraints it must satisfy. We also generalize the definition of observability to take time into consideration. We then prove that a supervisor exists if and only if the language to be synthesized is controllable, observable, and $L_m(G)$ -closed. If the language to be synthesized is not controllable and observable, then the problem of synthesizing an optimal supervisor becomes complicated, partly because, as to be shown in Section IV, the supremal controllable and observable sublanguage of a given maximal legal language need not exist. Two methods will be proposed to solve the synthesis problem. In the first method, a stronger version of observability, called normality, is defined. The supremal controllable and normal sublanguage of a given language exists, and preserves the $L_m(G)$ -closedness of the language, and hence a supervisor can be synthesized

Manuscript received June 22, 1994. This research is support in part by the National Science Foundation of USA under Grant ECS-9213922 and by the National Science and Engineering Research Council of Canada under Grant A-7399.

F. Lin is with the Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI 48202 USA.

W. M. Wonham is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario M5S 1A4, Canada.

IEEE Log Number 9407004.

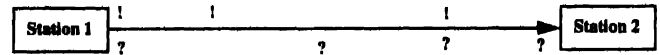


Fig. 1. (l-traffic light, ?-detector).

based on it. We also show that this stronger version of observability is actually equivalent to observability itself if all controllable (or prohibitable) events as well as *tick* are observable. In the second method a supervisor is first synthesized assuming full observation and then modified when some events become unobservable. We show that, under some conditions, the second method is better than the first in the sense that the second method yields a supervisor generating a larger language and it can be implemented more efficiently. Finally, an example of two trains is given to illustrate the results of the paper. The development in this paper is parallel to the corresponding development by the authors on untimed discrete-event systems [8].

II. PRELIMINARIES

In this section, we briefly review the relevant results of [1] on timed discrete-event systems. Following the notation and definitions of [1], a timed discrete-event system is modeled by an automaton

$$G = (\Sigma, Q, \delta, q_0, Q_m)$$

where the events are partitioned into 1) prospective events Σ_{pr} that have finite upper time bounds, 2) remote events Σ_{rm} whose upper time bounds are infinite, and 3) tick of the global clock:

$$\Sigma = \Sigma_{\text{pr}} \cup \Sigma_{\text{rm}} \cup \{\text{tick}\}.$$

Remark 1 To exclude the physically unrealistic possibility that events in $\Sigma - \{\text{tick}\}$ occur infinitely often during one unit of time, we require that G be activity loop free [1], that is,

$$(\forall q \in Q)(\forall s \in (\Sigma - \{\text{tick}\})^+) \delta(s, q) \neq q.$$

Remark 2 If G is constructed from the activity transition graph G_{act} and time bounds on events as shown in [1], then the advance of time will never stop. Under the assumption of activity loop freedom, this is equivalent to saying that G is type-1 blocking free [11], that is,

$$(\forall s \in L(G)) \Sigma_{L(G)}(s) \neq \emptyset$$

where $\Sigma_L(s) = \{\sigma \in \Sigma : s\sigma \in L\}$.

Remark 3 Again, if G is constructed from G_{act} and time bounds on events, then the fact that no *tick* is possible after a string implies that some prospective events must be possible after that string, that is,

$$(\forall s \in L(G)) \text{tick} \notin \Sigma_{L(G)}(s) \Rightarrow \Sigma_{\text{pr}} \cap \Sigma_{L(G)}(s) \neq \emptyset.$$

In this paper, we assume that G is constructed from G_{act} and time bounds on events.

Example 1 Consider a single track linking two stations as depicted in Fig. 1. To control the movement of trains, three traffic lights are installed along the track. Also installed are four detectors to observe the movement of trains. The track is divided accordingly into four sections: S_1 , S_2 , S_3 , and S_4 .

Suppose there are two trains T_1 and T_2 scheduled to depart from Station 1 to Station 2. For T_i , $i = 1, 2$ the designated events are as follows:

- σ_{11} : T_1 entering S_1 , σ_{14} : T_1 exiting S_4
- σ_{13} : T_1 entering S_2 , σ_{16} : T_1 exiting S_2
- σ_{12} : T_1 entering S_3 , σ_{18} : T_1 exiting S_3
- σ_{15} : T_1 entering S_4 , σ_{10} : T_1 exiting S_4

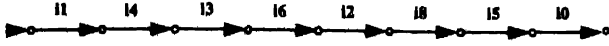


Fig. 2 (Numbers denote subscripts of event)

The activity transition graphs G_{act}^i (for T_i) is given in Fig. 2. The lower time bound l_σ and upper time bound u_σ for events σ are

σ	σ_{11}	σ_{13}	σ_{12}	σ_{15}	σ_{14}	σ_{16}	σ_{18}	σ_{10}
l_σ	0	0	0	0	1	1	2	1
u_σ	∞	∞	∞	∞	1	2	3	1

Hence the prospective events are

$$\Sigma_{pro} = \{\sigma_{14}, \sigma_{16}, \sigma_{18}, \sigma_{10}, \sigma_{21}, \sigma_{26}, \sigma_{28}, \sigma_{20}\}$$

and the remote events are

$$\Sigma_{rem} = \{\sigma_{11}, \sigma_{13}, \sigma_{12}, \sigma_{15}, \sigma_{21}, \sigma_{24}, \sigma_{22}, \sigma_{27}\}$$

Simple time bounds are used here for the sole purpose of limiting the number of states. They do, however, illustrate the advantages of the timed model (cf. Section VI). More general time bounds could be used at the expense of more computation.

From G_{act}^1 , G_{act}^2 , and the time bounds, we can construct the timed discrete event system G describing the concurrent movement of T_1 and T_2 , which has 256 states.

To introduce control, let $\Sigma_{hib} \subseteq \Sigma_{rem}$ be the set of prohibitable (or uncontrollable) events, $\Sigma_f \subseteq \Sigma_{rem} \cup \Sigma_{pro}$ be the set of forcible events, and $\Sigma_{uc} = (\Sigma_{rem} \cup \Sigma_{pro}) - \Sigma_{hib}$ be the set of uncontrollable events. The following definition [1] generalizes the definition in [13].

Definition 1 Let $K \subseteq L(G)$. We define K to be controllable (with respect to $L(G)$) if, for all $s \in \bar{K}$

- 1) $\Sigma_K(s) \cap \Sigma_f = \emptyset \Rightarrow \Sigma_{L(G)}(s) \cap (\Sigma \cup \{tick\}) \subseteq \Sigma_K(s)$ and
- 2) $\Sigma_K(s) \cap \Sigma_f \neq \emptyset \Rightarrow \Sigma_{L(G)}(s) \cap \Sigma \subseteq \Sigma_K(s)$

In other words, given history $s \in \bar{K}$, if there are no forcible events available then $tick$ is not controllable; if forcible events are available, then $tick$ can be postponed. In proofs we will often use the following equivalent definition for controllability:

- 1) $(\forall s \in \bar{K}) \Sigma_{L(G)}(s) \cap \Sigma_f \subseteq \Sigma_K(s)$ and
- 2) $(\forall s \in \bar{K}) \Sigma_K(s) \cap \Sigma_f = \emptyset \wedge tick \in \Sigma_{L(G)}(s) \Rightarrow tick \in \Sigma_K(s)$

III. OBSERVABILITY

In practice, a supervisor may not be able to detect the occurrences of all events. When that happens, the existence of a supervisor is no longer guaranteed by controllability alone; we need to introduce the concept of observability. Observability for untimed discrete-event systems was introduced in [3], [8]. In this paper, we generalize this concept to timed discrete event systems.

To this end, we denote the observable events by $\Sigma \subseteq \Sigma$ and the unobservable events by $\Sigma_u = \Sigma - \Sigma$. A projection $P: \Sigma^* \rightarrow \Sigma_u^*$ is defined inductively as follows:

$$P\epsilon = \epsilon, \quad \epsilon \text{ is the empty string}$$

$$P(s\sigma) = \begin{cases} P s, & \text{if } \sigma \in \Sigma \\ (P s)\sigma, & \text{if } \sigma \in \Sigma_u \end{cases}$$

In other words, if a sequence of events s occurred in G , a supervisor can only observe $P s$. Therefore, a supervisor is now described by $\text{map } \gamma: PL(G) \rightarrow 2^\Sigma$. The language generated by γ ($L(G, \gamma)$ called closed behavior), is defined inductively as follows:

- 1) $\epsilon \in L(G, \gamma)$,
- 2) If $s \in L(G, \gamma)$, $\sigma \in \gamma(P s)$, and $s\sigma \in L(G)$, then $s\sigma \in L(G, \gamma)$, and
- 3) No other strings belong to $L(G, \gamma)$.

The language marked by γ (called marked behavior) is defined as

$$L_m(G, \gamma) = L(G, \gamma) \cap L_m(G).$$

The supervisor γ is said to be nonblocking if

$$\overline{L_m(G, \gamma)} = L(G, \gamma)$$

We say a supervisor is admissible if it satisfies the following two requirements. First, no uncontrollable events can be disabled, and second, if $tick$ is physically possible and no forcible events can preempt it, then it cannot be disabled. Formally, for all $s \in L(G, \gamma)$, we require

- 1) $\Sigma_u \subseteq \gamma(P s)$, and
- 2) $\Sigma_{L(G)}(s) \cap \gamma(P s) \cap \Sigma_f = \emptyset \wedge tick \in \Sigma_{L(G)}(s) \Rightarrow tick \in \gamma(P s)$

The system supervised under an admissible supervisor cannot have type 1 blocking as noted in the following proposition.

Proposition 1 ([12]) For all $s \in L(G, \gamma)$

$$\Sigma_{L(G)}(s) \cap \gamma(P s) \neq \emptyset \quad \square$$

We show that for a supervisor to make a correct decision (i.e., to synthesize a language K) based on partial observation, the following observability condition must be satisfied.

Definition 2 Let $K \subseteq L(G)$. We define K to be observable (with respect to $L(G)$) if, for all $s, s' \in \Sigma^*$

$$P s = P s' \Rightarrow \text{consis}(s, s')$$

where $\text{consis}(s, s')$ is true if and only if

- 1) $(\forall \sigma \in \Sigma) s\sigma \in \bar{K} \wedge s' \in \bar{K} \wedge s'\sigma \in L(G) \Rightarrow s'\sigma \in \bar{K}$, and
- 2) $(\forall \sigma \in \Sigma) s'\sigma \in \bar{K} \wedge s \in \bar{K} \wedge s\sigma \in L(G) \Rightarrow s\sigma \in \bar{K}$

In words, two strings s and s' satisfy $\text{consis}(s, s')$ if and only if s and s' are consistent with respect to one-step continuations in K . Observability requires that if two strings look the same to a supervisor, then they must be consistent. This definition is a modification of observability as defined in [8].

This observability condition together with the controllability condition guarantees the existence of a supervisor as shown in the following theorem.

Theorem 1 ([12]) Let $K \subseteq L(G)$ be a nonempty language. There exists a nonblocking supervisor γ such that $L_m(G, \gamma) = K$ if and only if the following three conditions are all satisfied:

- 1) K is controllable with respect to $L(G)$
- 2) K is observable with respect to $L(G)$, and
- 3) K is $L(G)$ -closed.

If we are only interested in the closed behavior, then we have the following corollary.

Corollary 1 Let $K \subseteq L(G)$ be a nonempty language. There exists a supervisor γ such that $L(G, \gamma) = K$ if and only if the following three conditions are all satisfied:

- 1) K is controllable with respect to $L(G)$
- 2) K is observable with respect to $L(G)$, and
- 3) K is closed.

IV. NORMALITY

The language describing a control objective is not very likely to be controllable and observable when a control objective is first specified because the objective is usually laid down independently of the consideration of controllability and observability of events. When this happens, the original control objective is not achievable by a supervisor. So we will try to achieve the largest possible part of the original control objective. This means finding, if possible, the largest sublanguage of the original language that is controllable and

observable. Such a supervisor is "optimal" in the sense that a system generating a larger language can "run" faster when driven by the same stochastic process generating event lifetimes, if a certain "fairness" condition is satisfied. The reader is referred to [7] for this result.

Let $\emptyset \neq E \subseteq L_m(G)$ be the language describing the original control objective, called the maximal legal language. We assume that E is $L_m(G)$ -closed.

The supremal controllable and observable sublanguage of E does not exist in general (see [12, example 2]) because the set of observable languages is not closed under union. To overcome this difficulty, we first modify the definition of normality introduced in [8].

Definition 3: Let $K \subseteq L(G)$. We define K to be normal (with respect to $L(G)$) if, for all $s \in \Sigma^*$,

$$s \in L(G) \wedge Ps \in P\bar{K} \Rightarrow s \in \bar{K}. \quad \square$$

Since $s \in \bar{K} \Rightarrow s \in L(G) \wedge Ps \in P\bar{K}$ is always true, K is normal if and only if $L(G) \cap P^{-1}P\bar{K} = \bar{K}$. Therefore, if K is normal, we can check whether a string $s \in L(G)$ is in \bar{K} by checking whether its projection Ps is in $P\bar{K}$. In other words, information on occurrences of unobservable events is not needed in deciding whether $s \in \bar{K}$. Hence, we expect that normality will be stronger than observability.

Proposition 2 ([12]): If K is normal, the K is observable.

The set of normal languages is algebraically better behaved than that of observable languages, in the sense that it is closed under arbitrary unions.

Theorem 2 ([12]): The set

$$CN(E) = \{K \subseteq E: K \text{ is controllable and normal}\}$$

is nonempty and closed under arbitrary unions. Therefore, the supremal element of $CN(E)$, $\sup CN(E)$, exists and belongs to $CN(E)$.

Assume that all events are observable. Then normality is automatically satisfied. So we have the following corollary.

Corollary 2: The set

$$C(E) = \{K \subseteq E: K \text{ is controllable}\}$$

is nonempty and closed under arbitrary unions. Therefore, the supremal element of $C(E)$, $\sup C(E)$, exists and belongs to $C(E)$.

Similarly, by assuming that all events are controllable, we have the following corollary.

Corollary 3: The set

$$N(E) = \{K \subseteq E: K \text{ is normal}\}$$

is nonempty and closed under arbitrary unions. Therefore, the supremal element of $N(E)$, $\sup N(E)$, exists and belongs to $N(E)$.

From the above results, we propose a direct approach to synthesize a supervisor when E is not controllable and observable. The approach is to synthesize a supervisor for $\sup C'N(E)$. For this to work, however, we need to know that $\sup C'N(E)$ is also $L_m(G)$ -closed.

Proposition 3 ([12]): If E is $L_m(G)$ -closed, then $\sup C'N(E)$ is also $L_m(G)$ -closed.

Proposition 2 shows that normality is stronger than observability. But, how strong is it? The following proposition partially answers this question. It states that if *tick* and all controllable events are observable, then normality is equivalent to observability in the presence of controllability.

Proposition 4 ([12]): Assume $\Sigma_{\text{tick}} \cup \{\text{tick}\} \subseteq \Sigma_o$. If K is controllable and observable, then K is controllable and normal.

V. MODIFICATION FOR PARTIAL OBSERVATION

In the previous section, a direct approach was proposed to synthesize a supervisor if E is not controllable and observable, that is, a supervisor was synthesized for $\sup C'N(E)$. In this section, we

propose a different approach: we will first synthesize a supervisor for $\sup C(E)$ under the assumption of full observation and then modify the supervisor for partial observation. We show that, under certain conditions, this indirect approach leads to a supervisor with a larger closed behavior. Furthermore, as shown in [4], when the resulting supervisor is implemented on-line, the computational complexity for updating control after observing a new event is linear with respect to the number of states in the supervisor for full observation. In other words, in the worst case, the computational complexity of the indirect approach at each step is of the order $O(|X|)$, in comparison with the total complexity of $O(2^{|X|})$ for the direct approach, where $|X|$ is the number of states in the refined generator of K .

We need the following preliminary results.

Proposition 5 ([12]): The set

$$\overline{CN}(\bar{E}) = \{K \subseteq \bar{E}: K \text{ is closed, controllable, and normal}\}$$

is nonempty and closed under arbitrary unions. Therefore, the supremal element of $\overline{CN}(\bar{E})$ exists, belongs to $\overline{CN}(\bar{E})$, and is given by

$$\sup \overline{CN}(\bar{E}) = \sup C'N(\bar{E}). \quad \square$$

By this proposition, we do not need to distinguish $\sup \overline{CN}(\bar{E})$ and $\sup C'N(\bar{E})$.

Corollary 4: The set

$$\bar{C}(\bar{E}) = \{K \subseteq \bar{E}: K \text{ is closed and controllable}\}$$

is nonempty and closed under arbitrary unions. Therefore, the supremal element of $\bar{C}(\bar{E})$ exists, belongs to $\bar{C}(\bar{E})$, and is given by

$$\sup \bar{C}(\bar{E}) = \sup \bar{C}(\bar{E}). \quad \square$$

Corollary 5: The set

$$\bar{N}(\bar{E}) = \{K \subseteq \bar{E}: K \text{ is closed and normal}\}$$

is nonempty and closed under arbitrary unions. Therefore, the supremal element of $\bar{N}(\bar{E})$ exists, belongs to $\bar{N}(\bar{E})$, and is given by

$$\sup \bar{N}(\bar{E}) = \sup \bar{N}(\bar{E}). \quad \square$$

Since $\sup C'(\bar{E}) = \sup \bar{C}(\bar{E})$ is closed and controllable, we can synthesize a supervisor with full observation $\gamma': L(G) \rightarrow 2^X$ such that

$$L(G, \gamma') = \sup C'(\bar{E})$$

where $L(G, \gamma')$ is defined in the same way for $L(G, \gamma)$ except that $Ps = s$.

Under partial observation, we modify γ' to a supervisor $\gamma: PL(G) \rightarrow 2^X$ as follows. Let $[t] = P^{-1}(t) \cap L(G, \gamma')$ be the set of strings in $L(G, \gamma')$ having the projection t and

$$\gamma(t) = \bigcup_{s \in [t]} (\gamma'(s) \cup (\Sigma - \Sigma_{I(t)}(s))).$$

A similar modification was introduced in [4] for untimed discrete event systems. However, for a timed discrete-event system, the issues of forcible events and nonblocking need to be dealt with differently. To this end, we define coherence as follows.

Definition 4: Let $K \subseteq L(G)$. We define K to be coherent (with respect to $L(G)$) if, for all $s \in \bar{K}$,

$$\begin{aligned} \text{tick} \in \Sigma_{L(t)}(s) &\Rightarrow (\forall s', s'' \in P^{-1}Ps \cap \bar{K}) \\ &\quad (\Sigma_K(s') \cap \Sigma_{\text{tick}} = \Sigma_K(s'') \cap \Sigma_{\text{tick}}) \\ &\quad \wedge (\Sigma_{L(t)}(s') \cap \Sigma_{\text{tick}} = \Sigma_{L(t)}(s'') \cap \Sigma_{\text{tick}}). \end{aligned}$$

In words, if K is coherent, then for every string s that can be followed by *tick* in $L(G)$, the set of forcible events that are feasible (or legal) after any string having the same projection as s is the same.

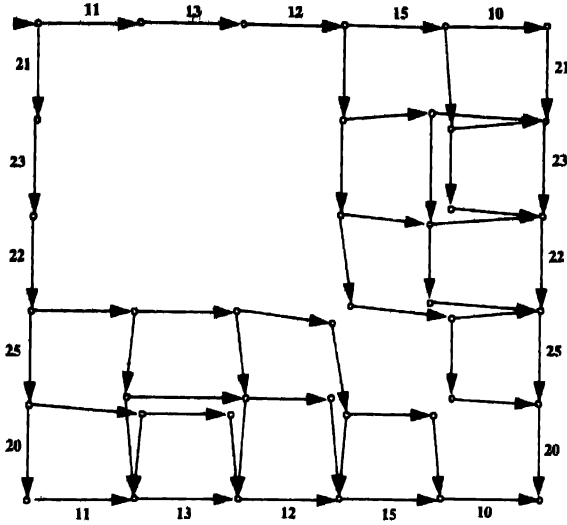


Fig. 3. (Horizontal events in same column, or vertical events in same row are same.)

Proposition 6 ([12]): If γ' is an admissible supervisor and $L(G, \gamma')$ is coherent, then γ is also an admissible supervisor.

This modified supervisor γ may block. To guarantee nonblocking, it is convenient to introduce the additional condition of livelock-freedom [5]. First, let us recall the definition of livelock-freedom given in [5].

Definition 5: Let $K \subseteq \Sigma^*$. We define K to be livelock free if, for all $s \in \bar{K}$

$$(\exists n \in \mathbb{N})(\forall t \in \Sigma^*)|t| \geq n \wedge st \in \bar{K} \Rightarrow (\exists u \in \Sigma^*) (s \leq u \leq st \wedge u \in K)$$

where \mathbb{N} denotes the set of natural numbers and $s \leq u$ denotes that s is a prefix of u .

In words, if K is livelock-free, then every infinite chain of strings in \bar{K} visits the set of marker states infinitely often.

Proposition 7 ([12]): If $L_m(G)$ is livelock-free and γ is admissible, then γ is nonblocking.

The implication of the above proposition is that, under the assumption of livelock-freedom, any admissible supervisor is nonblocking. Finally, we can prove that the modified supervisor generates only legal behavior that contains at least $\sup CN(E)$.

Theorem 3 ([12]): The language marked by the modified supervisor γ is bounded by

$$\sup CN(E) \subseteq L_m(G, \gamma) \subseteq E. \quad \square$$

To conclude this section, we note that $L(G, \gamma)$ obtained this way may not be maximal (see Example 3 in [12]).

VI. EXAMPLE

We continue our discussion of the example in Section II. Our control objective is to prevent two trains from colliding. So the corresponding maximal legal language E can be obtained by deleting states in G corresponding to two trains being in the same section at the same time.

Because of the locations of the traffic lights and detectors, the controllable and observable events are

$$\begin{aligned} \Sigma_{lib} &= \{\sigma_{11}, \sigma_{13}, \sigma_{15}, \sigma_{21}, \sigma_{23}, \sigma_{25}\} \\ \Sigma_o &= \{\sigma_{11}, \sigma_{12}, \sigma_{15}, \sigma_{10}, \sigma_{21}, \sigma_{22}, \sigma_{25}, \sigma_{20}\}. \end{aligned}$$

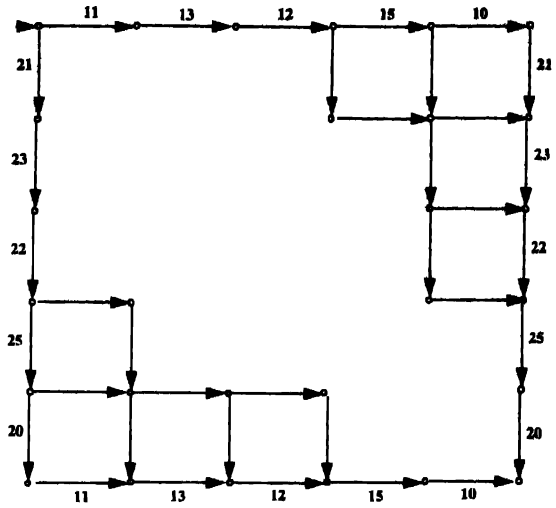


Fig. 4. (Horizontal events in same column, or vertical events in same row are the same.)

Since the movement of a train can be initiated, we assume the following events are forcible:

$$\Sigma_{for} = \{\sigma_{11}, \sigma_{13}, \sigma_{12}, \sigma_{15}, \sigma_{21}, \sigma_{23}, \sigma_{22}, \sigma_{25}\}.$$

To compare the result with that of untimed model, we use the direct method to synthesize a supervisor. We first calculate $\sup CN(E)$ and then synthesize a supervisor based on $\sup CN(E)$.

In [6], the same example is discussed in the untimed framework. Only events $\Sigma_{unt} = \{\sigma_{11}, \sigma_{13}, \sigma_{12}, \sigma_{15}, \sigma_{10}, \sigma_{21}, \sigma_{23}, \sigma_{22}, \sigma_{25}, \sigma_{20}\}$ are used in the untimed model. Therefore, we project $\sup CN(E)$ onto Σ_{unt} and compare it (Fig. 3) with the controlled behavior of the untimed model (Fig. 4) from [6]. Clearly, the timed model allows more strings to be generated. This is due to the additional information provided by the timed model. For example, in the untimed model, the fact that a train has left S_1 cannot be verified until it enters S_3 (where a detector is located). However, in the timed model, this can be inferred by the passage of one unit of time as a consequence of the time bounds imposed.

VII. CONCLUSION

We have generalized supervisory control under partial observation to timed discrete-event systems. Observability and normality are generalized to express the existence conditions of a supervisor for timed discrete event systems. Two approaches for supervisor synthesis are proposed. Although the detailed derivations of the results are different from those for untimed discrete-event systems, the two developments are parallel. Future research will focus on efficient implementations [2] and decentralized supervision [3], [8], [9].

REFERENCES

- [1] B. A. Brandin and W. M. Wonham, "Supervisory control of timed discrete-event systems," *IEEE Trans. Automat. Contr.*, vol. 39, no. 2, pp. 329–342, 1994.
- [2] S.-L. Chung, S. Lafortune, and F. Lin, "Limited lookahead policies in supervisory control of discrete-event systems," *IEEE Trans. Automat. Contr.*, vol. 37, no. 12, pp. 1921–1935, 1991.
- [3] R. Cieslak, C. Desclaux, A. Fawaz, and P. Varaiya, "Supervisory control of discrete-event processes with partial observations," *IEEE Trans. Automat. Contr.*, vol. 33, no. 3, pp. 249–260, 1988.
- [4] M. Heymann and F. Lin, "On-line control of partially observed discrete-event systems," Technion, Haifa, Israel, CIS Rep. 9310.
- [5] S. Lafortune and F. Lin, "On tolerable and desirable behaviors in supervisory control of discrete-event systems," *Discrete-Event Dynamic Systems: Theory and Applications*, vol. 1, no. 1, pp. 61–92, 1991.

- [6] F. Lin, "On controllability and observability of discrete-event systems," University of Toronto, Toronto, Ont., Canada, Ph.D. dissertation, 1987.
- [7] —, "Analysis of temporal performance of supervised discrete-event systems," *Automatica*, vol. 30, no. 3, pp. 533–536, 1994.
- [8] F. Lin and W. M. Wonham, "On observability of discrete-event systems," *Inform. Sci.*, vol. 44, no. 3, pp. 173–198, 1988.
- [9] —, "Decentralized supervisory control of discrete-event systems," *Inform. Sci.*, vol. 44, no. 3, pp. 199–224, 1988.
- [10] —, "Decentralized control and coordination of discrete-event systems with partial observation," *IEEE Trans. Automat. Contr.*, vol. 35, no. 12, pp. 1330–1337, 1990.
- [11] —, "Verification of nonblocking in decentralized supervision," *Control Theory and Advanced Technology*, vol. 7, no. 1, pp. 19–29, 1991.
- [12] —, "Supervisory control of timed discrete event systems under partial observation," University of Toronto, Toronto, Ont., Canada, Systems Control Group Rep. 9316, 1993.
- [13] P. J. Ramadge and W. M. Wonham, "Supervisory control of a class of discrete-event processes," *SIAM J. Contr. Optimization*, vol. 25, no. 1, pp. 206–230, 1987.
- [14] W. M. Wonham and P. J. Ramadge, "On the supremal controllable sublanguage of a given language," *SIAM J. Contr. Optimization*, vol. 25, no. 3, pp. 635–659, 1987.

Apology and Correction to "Process Control and Machine Learning: Rule-Based Incremental Control"

Dominique Luzeaux and Bertrand Zavidovique

APOLOGY AND CORRECTION

It has come to our attention that the name of the second author was omitted from the above paper appearing in the June, 1994, issue of *TRANSACTIONS ON AUTOMATIC CONTROL*, pp. 1166–1171. Due to an apparent error in transcription, a former editorial assistant omitted the name of Bertrand Zavidovique, who was the second author on this paper. Through a regrettable and remarkably coincidental sequence of oversights, the original error was not corrected either in the typesetting of the article or in the proofreading of the galleys by the first author. The *TRANSACTIONS* offers a sincere apology to Prof. Zavidovique for this mistake and hereby provides publication of the title and authors as it should have appeared in June. The paper in question appears in vol. 39, no. 6.

Book Reviews.

In this section, the IEEE Control Systems Society (CSS) publishes reviews of books in the control field and related areas. Readers are invited to send comments on these reviews for possible publication in the Technical Notes and Correspondence section of this TRANSACTIONS. The CSS does not necessarily endorse the opinions of the reviewers.

If you have used an interesting book for a course or as a personal reference, we encourage you to share your insights with our readers by writing a review for possible publication in the TRANSACTIONS. All material published in the TRANSACTIONS is reviewed prior to publication. Submit a completed review or a proposal for a review to:

D. S. Naidu
Associate Editor—Book Reviews
College of Engineering
Idaho State University
833 South Eighth Street
Pocatello, ID 83209

Mathematical Control Theory: Deterministic Finite-Dimensional Systems—Eduardo D. Sontag (New York: Springer-Verlag, 1990). *Reviewed by Stephen P. Boyd.*

The title of this book gives a very good description of its contents and style, although "Introduction to" could be added to the beginning. The style is mathematical: precise, clear statements (i.e., theorems) are asserted, then carefully proved. The book covers many of the key topics in control theory, except (as the subtitle has warned us) those involving stochastic processes or infinite-dimensional systems. The level is appropriate for a senior undergraduate majoring in mathematics or a graduate student in electrical engineering who has been exposed to the style of mathematics (in a first course on analysis or algebra). This book fills an important niche and can play a key role in increasing the awareness and appreciation of control theory among mathematicians.

The first chapter is a nicely written extended discussion about what control theory is. Sontag introduces some of the key concepts (e.g., system, feedback, state-space, and nonlinearity) with simple examples in an informal style. For an engineering graduate student, it will be a review of some undergraduate courses; for the mathematics student, it provides some cultural background for the rest of the book. This chapter is not in the same style as the remainder of the book so I might have numbered it Chapter 0. There are no theorems or proofs, and Sontag keeps the discussion at the level of informal ideas, for example, (p. 5) "Designs based on linearizations work locally for the original system." Pity the mathematics student who then searches backward for the precise definition of the terms "design" and "work"!

The book proper starts at Chapter 2, which covers the abstract idea of a system. The style jumps from the informality of Chapter 1 to a more in-depth style with a list of long mathematical definitions. Sontag writes that the chapter contains some "abstract nonsense" (p. ix), e.g., very general and abstract definitions of a dynamical system. On this topic he points out a very good analogy from the foundations of mathematics. A student of mathematics first thinks of a function in an informal way, as a "mapping" or "subroutine." The student then encounters the formal definition of a function from A to B as a subset of $A \times B$ which satisfies certain properties. Not long thereafter

the student reverts to the more informal model of a function as a "mapping" or "subroutine." The difference is that now the student really knows what a function is. In a similar way, Chapter 2 introduces the student to the formal definition of a dynamical system. After this introduction, the student can revert to a less formal idea of a dynamical system (e.g., "the state summarizes the effects of the past inputs on the future outputs"). But now, these ideas rest on a firm footing. Sontag punctuates the heavy definitions with nice examples, which makes the medicine of a formal foundation easier to take.

Chapter 3 covers various topics involving reachability and controllability. Sontag examines these topics in some depth for the important special case of linear, time-invariant (LTI) systems. The treatment might be too brief for the student who has had no previous exposure to LTI systems; this is no problem for the engineering graduate student, who will have had one or possibly two courses which use these ideas.

Chapter 4 covers (state) feedback, with a bit on Lyapunov stability. In this chapter, the author gives a precise statement of the linearization principle described informally in Chapter 1. Chapter 5 covers the dual notion of observability as well as an introduction to realization theory and minimality. Although Section 5.8, Abstract Realization Theory,* is marked as skippable, it shows the advantage of the abstract approach: generality. As an example, Sontag constructs a minimal realization of a parity check system.

In Chapter 6 Sontag considers observers and dynamic feedback. He cannot give a detailed discussion of the design of observers (e.g., the trade-off between convergence rate and noise sensitivity) since he assumes no knowledge of stochastic processes, but he does give a deterministic version of the Kalman filter in the next (and final) chapter, which covers optimal control. Two sections in Chapter 6 are too brief for students who have not been introduced to the topics: frequency-domain considerations (Section 6.4) and parameterization of stabilizers (Section 6.5). These sections should have been marked as optional or even not covered (like stochastic processes).

Chapter 7 covers optimal control, taking Bellman's dynamic programming approach. One of the advantages of this approach is that it applies to more abstract systems (e.g., finite state systems) so an example of this type would be useful before launching into the traditional case of a linear system with quadratic cost. This chapter ends with a nice deterministic development of the Kalman filter. Here the goal is to find the initial state that minimizes an integral quadratic measure of the difference between the output observed and the output predicted, given the initial state.

The reviewer is with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA.
IEEE Log Number 9407567.

The appendixes cover very useful material, e.g., singular value decomposition. It is a pity that some of these topics are not covered in the traditional mathematics curriculum.

The book contains about 400 references. A list of references does not have great added value today because of computerized literature searches. Sontag, however, ends each chapter with a very useful Notes and Comments section with annotated pointers to the references. The references seem to be concentrated in Western journals (e.g., IEEE TRANSACTIONS ON AUTOMATIC CONTROL, *Systems and Control Letters*); only a handful are from the former Soviet Union, which has a very strong tradition in mathematical control theory tracing back to Pontriagin and Lyapunov. I must admit that this comment applies equally as well to the books I have written.

There are several uses for this book. It would work very well as the text for an undergraduate senior-level mathematics course. This course could replace, or at least complement, the traditional course on 19th century calculus of variations that is often found as a senior-level mathematics elective. The course could also serve graduate students in engineering.

The second use for the book is as a reference or supplementary text for engineering graduate students, particularly Ph.D. students. These students have already been exposed to many of the ideas in Sontag's book in engineering-oriented courses. Often in their first year of Ph.D. study they firm up their mathematics background by taking mathematics courses such as analysis, functional analysis, abstract algebra, and differential geometry. Sontag's book serves as an excellent bridge between the two disciplines: it covers topics traditionally treated in engineering courses, but in a mathematical style.

One similar book is Vidyasagar [3], which is oriented more towards the advanced engineering student and I think more appropriate for an engineering course on advanced control theory. The book by Delchamps [1] could serve as a secondary text with Sontag's book, since it covers more linear systems theory, also in a precise mathematical style. The book by Wonham [4] covers linear system and control theory in the most elegant mathematical style. It could serve as the text for a sequel to the introductory mathematical control theory course based on Sontag's book. The recent text by Zabczyk [5], *Mathematical Control Theory: An Introduction*, is similar to Sontag's book in style, level, and coverage. Zabczyk's choice of more advanced topics differs from Sontag's; Zabczyk includes more on optimal control as well as a fairly complete introduction to infinite-dimensional systems.

Sontag clearly put much thought and effort into this book, and it shows. The book succeeds in conveying the important basic ideas of mathematical control theory, with appropriate level and style, to seniors in mathematics.

REFERENCES

- [1] D. F. Delchamps, *State-Space and Input-Output Linear Systems*. New York: Springer-Verlag, 1988.
- [2] E. D. Sontag, *Mathematical Control Theory*. New York: Springer-Verlag, 1990.
- [3] M. Vidyasagar, *Nonlinear Systems Analysis*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1992.
- [4] W. M. Wonham, *Linear Multivariable Control*. New York: Springer-Verlag, 1985.
- [5] J. Zabczyk, *Mathematical Control Theory: An Introduction*. Boston, MA: Birkhauser, 1992.

1994 Index

IEEE Transactions on Automatic Control

Vol. 39

This index covers all technical items — papers, correspondence, reviews, etc. — that appeared in this periodical during 1994, and items from previous years that were commented upon or corrected in 1994.

The Author Index contains the primary entry for each item, listed under the first author's name, and cross-references from all coauthors. The Subject Index contains several entries for each item under appropriate subject headings, and subject cross-references.

It is always necessary to refer to the primary entry in the Author Index for the exact title, coauthors, and comments/corrections.

AUTHOR INDEX

A

- Abdallah, C., see Perez, F., *T-AC Jul 94* 1470-1472
- Abed, E.H. Review of "Nonlinear System Analysis, 2nd edn." (Vidyasagar, M., 1993), *T-AC Jul 94* 1535-1536
- Abou-Kandil, H., G. Freiling, and G. Jank. Solution and asymptotic behavior of coupled Riccati equations in jump linear systems, *T-AC Aug 94* 1631-1636
- Afacan, T., and O. Yuksel. On the decomposition of Roesser's 2-D system model, *T-AC Nov 94* 2261-2262
- Aganovic, Z., and Z. Gajic. The successive approximation procedure for finite-time optimal control of bilinear systems, *T-AC Sep 94* 1932-1935
- Agathoklis, P., see Sreeram, V., *T-AC Feb 94* 381-385
- Agathoklis, P., see Sreeram, V., *T-AC May 94* 1102-1105
- Aggoune, M.E., F. Boudjema, A. Bensenuoui, A. Hellal, M.R. Elmesai, and S.V. Vadari. Design of variable structure voltage regulator using pole assignment technique, *T-AC Oct 94* 2106-2110
- Ahmed, M.S. A new algorithm for state estimation of stochastic linear discrete systems, *T-AC Aug 94* 1652-1656
- Ahmed, N.U., and S.M. Radaideh. Modified extended Kalman filtering, *T-AC Jun 94* 1322-1326
- Ailon, A., see Kelly, R., *T-AC Jun 94* 1222-1224
- Aitken, V.C., and H.M. Schwartz. On the exponential stability of discrete-time systems with applications in observer design, *T-AC Sep 94* 1959-1962
- Alberts, T.E., see Pota, H.R., *T-AC Aug 94* 1774-1776
- Aldeen, M., see Trinh, H., *T-AC Sep 94* 1948-1951
- Al-Muthairi, N.F., see Mahmoud, M.S., *T-AC May 94* 995-999
- Al-Muthairi, N.F., see Mahmoud, M.S., *T-AC Oct 94* 2135-2139
- Altman, E. Flow control using the theory of zero sum Markov games, *T-AC Apr 94* 814-818
- Anderson, B.D.O., see Limebeer, D.J.N., *T-AC Jan 94* 69-82
- Anderson, B.D.O., see Wei-Yong Yan, *T-AC Nov 94* 2347-2354
- Annaswamy, A.M., see Seto, D., *T-AC Jul 94* 1411-1428
- Annaswamy, A.M., see Karason, S.P., *T-AC Nov 94* 2325-2330
- Antonio, J.K., W.K. Tsai, and G.M. Huang. Time complexity of a path formulated optimal routing algorithm, *T-AC Feb 94* 385-391
- Antonio, J.K., W.K. Tsai, and G.M. Huang. Time complexity of a path formulated optimal routing algorithm (second printing), *T-AC Sep 94* 1839-1844
- Antsaklis, P.J., see Passino, K.M., *T-AC Feb 94* 269-279
- Antsaklis, P.J., see Passino, K.M., *T-AC Jul 94* 1531
- Antsaklis, P.J., see Schneider, A.M., *T-AC Feb 94* 435-441
- Aruga, S., see Miyamoto, S., *T-AC Feb 94* 379-381
- Aruga, M.A., see Yu Tang, *T-AC Sep 94* 1871-1875
- Arntis, K.G., see Paraskevopoulos, P.N., *T-AC Apr 94* 793-797
- Arm, K.J., C.C. Hang, and B.C. Lim. A new Smith predictor for controlling a process with an integrator and long dead-time, *T-AC Feb 94* 343-345
- Arns, M., see Douglas, J., *T-AC Jan 94* 107-111
- Balakrishnan, J., see Narendra, K.S., *T-AC Sep 94* 1861-1866
- Balakrishnan, J., see Narendra, K.S., *T-AC Dec 94* 2469-2471
- Balestrino, A., F. Bernini, and A. Landi. Steady-state behavior in the vibrational control of a class of nonlinear systems by AP-forcing, *T-AC Jun 94* 1255-1258
- Banavar, R.N., and J.L. Speyer. Risk-sensitive estimation and a differential game, *T-AC Sep 94* 1914-1918
- Banda, S.S., see Hsu, C.S., *T-AC Aug 94* 1679-1681
- Bandyopadhyay, B., see Pimpalkhare, A.A., *T-AC May 94* 1148
- Bandyopadhyay, B., O. Ismail, and R. Gorez. Routh-Pade approximation for interval systems, *T-AC Dec 94* 2454-2456
- Baras, J.S., see James, M.R., *T-AC Apr 94* 780-792
- Barghouti, I.S., see Schneider, A.M., *T-AC Feb 94* 435-441
- Baril, C., see Gutman, P.-O., *T-AC Jun 94* 1268-1273
- Barret, M., and M. Benidir. On the boundary of the set of Schur polynomials and applications to the stability of 1-D and 2-D digital recursive filters, *T-AC Nov 94* 2335-2339
- Basar, T., see Zigang Pan, *T-AC Feb 94* 280-299
- Bassong-Onana, A., see Darouach, M., *T-AC Aug 94* 1755-1758
- Baumgartner, E.T., and S.B. Skaar. An autonomous vision-based mobile robot, *T-AC Mar 94* 493-502
- Bayard, D.S. Extended horizon liftings for stable inversion of nonminimum-phase systems, *T-AC Jun 94* 1333-1338
- Bayard, D.S. An algorithm for state-space frequency domain identification without windowing distortions, *T-AC Sep 94* 1880-1885
- Beghi, A., A. Lepschy, and U. Viaro. A property of the Routh table and its use, *T-AC Dec 94* 2494-2496
- Bender, D.J. Comments on "On a property of compensators designed by the separation principle" [with reply], *T-AC Feb 94* 447-448
- Benidir, M., see Barret, M., *T-AC Nov 94* 2335-2339
- Bensenuoui, A., see Aggoune, M.E., *T-AC Oct 94* 2106-2110
- Bentsman, J., see Lehman, B., *T-AC May 94* 898-912
- Bentsman, J., see Keum Shik Hong, *T-AC Oct 94* 2018-2033
- Benveniste, A., see Chou, K.C., *T-AC Mar 94* 464-478
- Benzaid, Z., and M. Szaier. Constrained controllability of linear impulse differential systems, *T-AC May 94* 1064-1066
- Benzaouia, A. The resolution of equation $XA+XB=HX$ and the pole assignment problem, *T-AC Oct 94* 2091-2095
- Berghuis, H., and H. Nijmeijer. Robust control of robots via linear estimated state feedback, *T-AC Oct 94* 2159-2162
- Bernini, F., see Balestrino, A., *T-AC Jun 94* 1255-1258
- Bernstein, D.S., see Shuoh Ren, *T-AC Jan 94* 162-164
- Bernstein, D.S., see Haddad, W.M., *T-AC Jan 94* 229-234
- Bernstein, D.S., see Haddad, W.M., *T-AC Apr 94* 827-831
- Bernstein, D.S., and W.M. Haddad. Nonlinear controllers for positive real systems with arbitrary input nonlinearities, *T-AC Jul 94* 1513-1517
- Bernstein, D.S., see Haddad, W.M., *T-AC Aug 94* 1772
- Bernstein, D.S., see Wang, Y.W., *T-AC Nov 94* 2284-2287
- Bertrand, P., see Dufour, F., *T-AC Nov 94* 2354-2357
- Bettayeb, M., see Kavranoğlu, D., *T-AC Sep 94* 1899-1904
- Bhattacharyya, S.P., see Keel, L.H., *T-AC Jul 94* 1524-1530
- Bialasiewicz, J.T., see Radenkovic, M.S., *T-AC Feb 94* 396-400
- Bin Yao, S.P. Chan, and Danwei Wang. Unified formulation of variable structure control schemes for robot manipulators, *T-AC Feb 94* 371-376
- Birkholzer, T., see Kreisselmeier, G., *T-AC Jan 94* 33-46
- Bitmead, R.R., see Zhuquan Zang, *T-AC Jan 94* 171-175
- Bitmead, R.R., see Wei-Yong Yan, *T-AC Nov 94* 2347-2354
- Bittanti, S., and M. Campi. Bounded error identification of time-varying parameters by RLS techniques, *T-AC May 94* 1106-1110
- Blanchini, F. Ultimate boundedness control for uncertain discrete-time systems via set-induced Lyapunov functions, *T-AC Feb 94* 428-433
- Blanchini, F., and M. Szaier. Rational L^1 suboptimal compensators for continuous-time systems, *T-AC Jul 94* 1487-1492
- Bodenheimer, B., see Kemin Zhou, *T-AC Aug 94* 1564-1574
- Bodenheimer, B., see Doyle, J., *T-AC Aug 94* 1575-1587
- Bodson, M. Tuning, multitone instabilities, and intrinsic differences in robustness of adaptive control systems, *T-AC Apr 94* 864-870
- Bodson, M., see de Mathelin, M., *T-AC Aug 94* 1612-1617
- Bodson, M., A. Sacks, and P. Khosla. Harmonic generation in adaptive feedforward cancellation schemes, *T-AC Sep 94* 1939-1944
- Bolzern, P., P. Colaneri, and G. De Nicolao. On the computation of upper covariance bounds for perturbed linear systems, *T-AC Mar 94* 623-626

B

- Bouloul, J., see Seto, D., *T-AC Jul 94* 1411-1428
- Bouloul, J., see Seto, D., *T-AC Dec 94* 2442-2453
- Bouloglu, B., see Koc, C.K., *T-AC Aug 94* 1644-1647

- Bonivento, C., A. Nersisian, A. Tonielli, and R. Zanasl. A cascade structure for robust control design, *T-AC Apr 94* 846-849
- Bo Peng Rao, *see* Morgul, O., *T-AC Oct 94* 2140-2145
- Borisov, A.V., and A.R. Pankov. Optimal filtering in stochastic discrete-time systems with unknown inputs, *T-AC Dec 94* 2461-2464
- Bose, N.K. Argument conditions for Hurwitz and Schur polynomials from network theory, *T-AC Feb 94* 345-346
- Bose, S., A. Patra, and S. Mukhopadhyay. On observability with delay: antitheses and syntheses, *T-AC Apr 94* 803-806
- Bosov, A.V., *see* Pankov, A.R., *T-AC Aug 94* 1617-1620
- Bo Tang, Z. Adaptive partitioned random search to global optimization, *T-AC Nov 94* 2235-2244
- Boudjemaa, F., *see* Aggoune, M.E., *T-AC Oct 94* 2106-2110
- Bourlès, H. Semi-cancellable fractions in system theory, *T-AC Oct 94* 2148-2153
- Bouwels, J.P.H.M., *see* Damen, A.A.H., *T-AC May 94* 1075-1078
- Braatz, R.P., P.M. Young, J.C. Doyle, and M. Morari. Computational complexity of μ calculation, *T-AC May 94* 1000-1002
- Brandin, B.A., and W.M. Wonham. Supervisory control of timed discrete-event systems, *T-AC Feb 94* 329-342
- Bridges, M.M., *see* Grabbe, M.T., *T-AC Jan 94* 179
- Brogliato, B., and R. Lozano. Adaptive control of first-order nonlinear systems with reduced knowledge of the plant parameters, *T-AC Aug 94* 1764-1768
- Budka, K.C. Stochastic monotonicity and concavity properties of rate-based flow control mechanisms, *T-AC Mar 94* 544-548
- Bugong Xu, and Yongqing Liu. An improved Razumikhin-type theorem and its applications, *T-AC Apr 94* 839-841
- Bugong Xu. Comments on "Robust stability of delay dependence for linear uncertain systems", *T-AC Nov 94* 2365
- Bunse-Gerstner, A., V. Mehrmann, and N.K. Nichols. Regularization of descriptor systems by output feedback, *T-AC Aug 94* 1742-1748
- Burgat, C., *see* Tarbouriech, S., *T-AC Feb 94* 401-405
- Byrnes, C.I., and Wei Lin. Losslessness, feedback equivalence, and the global stabilization of discrete-time nonlinear systems, *T-AC Jan 94* 83-98
- Byrnes, C.I., *see* Wei Lin, *T-AC Nov 94* 2340-2346

C

- Campbell, S.L., E. Moore, and Yangchun Zhong. Utilization of automatic differentiation in control algorithms, *T-AC May 94* 1047-1052
- Campi, M., *see* Bittanti, S., *T-AC May 94* 1106-1110
- Campo, P.J., and M. Morari. Achievable closed-loop properties of systems under decentralized control: conditions involving the steady-state gain, *T-AC May 94* 932-943
- Caramanis, M., *see* Liberopoulos, G., *T-AC Apr 94* 889-895
- Carazo, A.H., and J.L. Perez. Linear estimation for discrete-time systems in the presence of time-correlated disturbances and uncertain observations, *T-AC Aug 94* 1636-1638
- Cassandras, C.G., *see* Sparagkis, P.D., *T-AC Jul 94* 1492-1497
- Castro-Linares, R., and C.H. Moog. Structure invariance for uncertain nonlinear systems, *T-AC Oct 94* 2154-2158
- Chakravorti, B. Optimal flow control of an M/M/1 queue with a balanced budget, *T-AC Sep 94* 1918-1921
- Chan, S.P., *see* Bin Yao, *T-AC Feb 94* 371-376
- Chang-Ik Kang, *see* In-Joong Ha, *T-AC Mar 94* 673-677
- Chang Yang, and P.T. Kabamba. Multi-channel output gain margin improvement using generalized sampled-data hold functions, *T-AC Mar 94* 657-661
- Changyun Wen. A robust adaptive controller with minimal modifications for discrete time-varying systems, *T-AC May 94* 987-991
- Changyun Wen. Decentralized adaptive regulation, *T-AC Oct 94* 2163-2166
- Chen, B.M., *see* Stoorvogel, A.A., *T-AC Feb 94* 355-360
- Chen, B.M., *see* Stoorvogel, A.A., *T-AC Sep 94* 1936-1939
- Chen, C.M., *see* Polyak, B.T., *T-AC May 94* 1147-1148
- Chen, T., *see* Qui, L., *T-AC Dec 94* 2506-2511
- Chen-Chung Liu, *see* Fu-Chuang Chen, *T-AC Jun 94* 1306-1310
- Cheng, T.C.E., and A. Janiak. Resource optimal control in some single-machine scheduling problems, *T-AC Jun 94* 1243-1246
- Cheng-Shang Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks, *T-AC May 94* 913-931
- Chia-Chi Tsui. A general failure detection, isolation and accommodation system with model uncertainty and measurement noise, *T-AC Nov 94* 2318-2321
- Chiaverini, S., B. Siciliano, and L. Villani. Force/position regulation of compliant robot manipulators, *T-AC Mar 94* 647-652
- Chih-Hai Fan, J.L. Speyer, and C.R. Jaensch. Centralized and decentralized solutions of the linear-exponential-Gaussian problem, *T-AC Oct 94* 1986-2003
- Ching-Fang Lin, *see* Jie Huang, *T-AC Jul 94* 1510-1513
- Ching-Fang Lin, *see* Jie Huang, *T-AC Nov 94* 2307-2311

D

- Chisel, L., and E. Mosca. Stabilizing I-O receding horizon control of CARMA plants, *T-AC Mar 94* 614-618
- Chockalingam, G., and S. Dasgupta. Strong stabilizability of systems with multiaffine uncertainties and numerator denominator coupling, *T-AC Sep 94* 1955-1958
- Chong, E.K.P., and P.J. Ramadge. Stochastic optimization of regenerative systems using infinitesimal perturbation analysis, *T-AC Jul 94* 1400-1410
- Chou, K.C., A.S. Willsky, and A. Benveniste. Multiscale recursive estimation, data fusion, and regularization, *T-AC Mar 94* 464-478
- Chou, K.C., A.S. Willsky, and R. Nikoukhan. Multiscale systems, Kalman filters, and Riccati equations, *T-AC Mar 94* 479-492
- Chow, J.H., *see* Date, R., *T-AC Feb 94* 347-351
- Chun-Hsiung Fang, and Li Lee. Robustness of regional pole placement for uncertain continuous-time implicit systems, *T-AC Nov 94* 2303-2307
- Chun-Yi Su, and Y. Stepanenko. Robust motion/force control of mechanical systems with classical nonholonomic constraints, *T-AC Mar 94* 609-614
- Chwan-Lu Tseng, *see* Juang-Huei Su, *T-AC Jun 94* 1341-1344
- Cishen Zhang, and R.J. Evans. Continuous direct adaptive control with saturation input constraint, *T-AC Aug 94* 1718-1722
- Clarke, D.W., E. Mosca, and R. Scattolini. Robustness of an adaptive predictive controller, *T-AC May 94* 1052-1056
- Cluett, W.R., *see* Wang, L., *T-AC Jul 94* 1463-1467
- Cobb, J.D. A unified theory of full-order and low-order observers based on singular system theory, *T-AC Dec 94* 2497-2502
- Colaneri, P., *see* Bolzern, P., *T-AC Mar 94* 623-626
- Collins, E.G., Jr., *see* Davis, I.D., *T-AC Apr 94* 849-852
- Collins, E.G., Jr., *see* Ge, Y., *T-AC Jun 94* 1302-1305
- Commercon, J.C. Eigenvalues of tridiagonal symmetric interval matrices, *T-AC Feb 94* 377-379
- Conrad, F., *see* Morgul, O., *T-AC Oct 94* 2140-2145
- Constantinescu, T., *see* Sayed, A.H., *T-AC May 94* 960-976
- Constantinescu, T., A.H. Sayed, and T. Kailath. A recursive Schur-based solution of the four-block problem, *T-AC Jul 94* 1476-1481
- Costa, O.L.V. Linear minimum mean square error estimation for discrete-time Markovian jump linear systems, *T-AC Aug 94* 1685-1689
- Costa, R.R., *see* Liu Hsu, *T-AC Jan 94* 4-21

- Dah-Ching Yang, *see* Wen-June Wang, *T-AC Jan 94* 99-102
- Dahleh, M.A., *see* Livstone, M.M., *T-AC Jul 94* 1531
- Dajun Wang, *see* Quan Wang, *T-AC Aug 94* 1711-1713
- d'Alessandro, P., and E. De Santis. Positiveness of dynamic systems with nonpositive coefficient matrices, *T-AC Jan 94* 131-134
- Damen, A.A.H., H.M. Falkus, and J.P.H.M. Bouwels. Modeling and control of a floating platform, *T-AC May 94* 1075-1078
- Danwei Wang, *see* Bin Yao, *T-AC Feb 94* 371-376
- Darouach, M., M. Zasadzinski, and S.J. Xu. Full-order observers for linear systems with unknown inputs, *T-AC Mar 94* 606-609
- Darouach, M. On the novel approach to the design of unknown input observers, *T-AC Mar 94* 698-699
- Darouach, M., M. Zasadzinski, and A. Bassong-Onana. Connection between the three-block generalized Riccati equation and the standard Riccati equation, *T-AC Aug 94* 1755-1758
- Dasgupta, S., *see* Garnett, J., *T-AC Jul 94* 1387-1399
- Dasgupta, S., *see* Chockalingam, G., *T-AC Sep 94* 1955-1958
- Date, R., and J.H. Chow. Decentralized stable factors and a parameterization of decentralized controllers, *T-AC Feb 94* 347-351
- Datta, A., and Ming-Tzu Ho. On modifying model reference adaptive control schemes for performance improvement, *T-AC Sep 94* 1977-1980
- Dattam, A., and P.A. Ioannou. Performance analysis and improvement in model reference adaptive control, *T-AC Dec 94* 2370-2387
- D'Attellis, C.E., *see* Gonzalez, G.A., *T-AC Oct 94* 2145-2148
- Davis, L.D., E.G. Collins, Jr., and A.S. Hodel. A parameterization of minimal plants, *T-AC Apr 94* 849-852
- Davis, L.D., *see* Ge, Y., *T-AC Jun 94* 1302-1305
- Dawson, D.M., *see* Zhihua Qu, *T-AC Nov 94* 2219-2234
- de Araujo, A.D., *see* Liu Hsu, *T-AC Jan 94* 4-21
- De Keyser, R., *see* Fanjin Kong, *T-AC Jul 94* 1467-1470
- de la Barra S., B.A.L. Correction to "On undershoot in SISO systems" *Mar 94* 578-581, *T-AC Aug 94* 1771
- de Mathelin, M., and M. Bodson. Multivariable adaptive control: identifiable parameterizations and parameter convergence, *T-AC Aug 94* 1612-1617
- Dembo, A., and O. Zeitouni. A large deviations analysis of range tracking loops, *T-AC Feb 94* 360-364
- Demetriou, M.A., and I.G. Rosen. On the persistence of excitation for adaptive estimation of distributed parameter systems, *T-AC Mar 94* 1117-1123
- De Nicolao, G., *see* Bolzern, P., *T-AC Mar 94* 623-626

- Derbel, N., M B A Kamoun, and M Poloujadoff New approach to block-diagonalization of singularly perturbed systems by Taylor expansion, *T-AC Jul 94* 1429-1431
- Derong Liu, see Kaining Wang, *T-AC Jun 94* 1251-1255
- De Santis, E., see d'Alessandro, P, *T-AC Jan 94* 131-134
- De Santis, E. On positively invariant sets for discrete-time linear systems with disturbance an application of maximal disturbance sets, *T-AC Jan 94* 245-249
- de Souza, C.E., see Lihua Xie, *T-AC Jun 94* 1310-1314
- Di Benedetto, M.D., A Glumineau, and C H Moog The nonlinear interactor and its application to input-output decoupling, *T-AC Jun 94* 1246-1250
- Di Benedetto, M.D., and J W Grizzle Asymptotic model matching for nonlinear systems, *T-AC Aug 94* 1539-1550
- Di Benedetto, M.D., see Grizzle, J W, *T-AC Sep 94* 1782-1794
- DiCesare, F., see Gius, A, *T-AC Apr 94* 818-823
- Diduch, C.P., see Marquez, H J, *T-AC Mar 94* 664-668
- Ding, X., L Guo, and P M Frank Parameterization of linear observers and its application to observer design, *T-AC Aug 94* 1648-1652
- Djafaris, T.E., see Kaminsky, R D, *T-AC Apr 94* 874-876
- Docampo, D., see Perez, F, *T-AC Jul 94* 1470-1472
- Dohyoung Chung, Taesam Kang, and Jang Gyu lee Stability robustness of LQ optimal regulators for the performance index with cross-product terms, *T-AC Aug 94* 1698-1702
- Dolphus, R.M. Sampled-data controller design for uncertain systems *T-AC May 94* 1036-1042
- Dong Xiang see Jian-Qiang Hu, *I-AC Mar 94* 640-643
- Dorsey, J.F., see Zhihua Qu *T-AC Nov 94* 2219-2234
- Douglas, J., and M Athans Robust linear quadratic designs with real parameter uncertainty, *T-AC Jan 94* 107-111
- Doyle, J., see Kemin Zhou, *T-AC Aug 94* 1564-1574
- Doyle, J., K Zhou, K Glover, and B Bodenheimer Mixed H_2 and H_∞ performance objectives II Optimal control *T-AC Aug 94* 1575-1587
- Doyle, J.C., see Braatz, R P, *T-AC May 94* 1000-1002
- Doyle, J.C., see Lu, W-M, *T-AC Dec 94* 2517-2524
- Duan, G.-R., and M-Z Wang Properties of the entire set of Hurwitz polynomials and stability analysis of polynomial families *T-AC Dec 94* 2490-2494
- Dufour, F., and P Bertrand Stabilizing control law for hybrid models, *T-AC Nov 94* 2354-2357
- Dugard, L see Gadjne, M, *T-AC Jun 94* 1259-1262
- Duksun Shim see Weiqian Sun *I-AC Oct 94* 2034-2046
- Duncan, T.E. P Mandl and B Pasik-Duncan On statistical sampling for system testing *T-AC Jan 94* 118-122
- Dupont, P.E. Avoiding stick-slip through PD control *T-AC May 94* 1094-1097
- Dwyer, I.A.W. see Karray F *T-AC May 94* 1016-1020
- Dye-Jyun Ma and A M Makowski On the convergence and ODE limit of a two-dimensional stochastic approximation *I-AC Jul 94* 1439-1442
- Dye-Jyun Ma, and Xi-Ren Ao A direct approach to decentralized control of service rates in a closed Jackson network *T-AC Jul 94* 1460-1463

E

- Egeland, O. and J-M Godhavn Passivity-based adaptive attitude control of a rigid spacecraft, *T-AC Apr 94* 842-846
- Egeland, O., and J-M Godhavn A note on Lyapunov stability for an adaptive robot controller, *T-AC Aug 94* 1671-1673
- Egeland, O., see Kanestrom R k *T-AC Sep 94* 1925-1928
- Eldem, V. The solution of diagonal decoupling problem by dynamic output feedback and constant precompensator the general case *T-AC Mar 94* 503-511
- Eldem, V., and H Selbuz On the general solution of the state deadbeat control problem, *T-AC May 94* 1002-1006
- El Erik, S., and R P Malhame Identification of alternating renewal electric load models from energy measurements, *T-AC Jun 94* 1184-1196
- Ellott, R.J., see James, M R, *T-AC Apr 94* 780-792
- Ellis, R.D., R Ravikanth, and P R Kumar Automating the simulation of complex discrete-time control systems a mathematical framework, algorithms, and a software package, *T-AC Sep 94* 1795-1801
- Elia, M.R., see Aggoune, M E, *T-AC Oct 94* 2106-2110
- Elverda, J.C. Calculation of an approximate solution of the infinite time-varying LQ-problem, *T-AC Jan 94* 235-238
- Elon, M.A., R S Smith, and A J Laub Finite-dimensional approximation and error bounds for spectral systems with partially known eigenstructure, *T-AC Sep 94* 1904-1909
- Elm, M. Comments on "Improved bounds for linear discrete-time systems with structured perturbations", *T-AC Aug 94* 1768-1769
- Elm, E.G., and R S S Pena Computation of algebraic combinations of uncertainty value sets, *T-AC Nov 94* 2315-2318
- Elm, R., see Musicki, D, *T-AC Jun 94* 1237-1241
- Elm, R.J., see Cishen Zhang, *T-AC Aug 94* 1718-1722

F

- Falkus, H.M., see Damen, A A H, *T-AC May 94* 1075-1078
- Fang, Y., K A Loparo, and X Feng Inequalities for the trace of matrix product, *T-AC Dec 94* 2489-2490
- Fanjun Kong, and R De Keyser Criteria for choosing the horizon in extended horizon predictive control, *T-AC Jul 94* 1467-1470
- Farrell, J.A., see Livstone, M M, *T-AC Jul 94* 1531
- Feng, X., see Fang, Y, *T-AC Dec 94* 2489-2490
- Feng Chun-Bo, see Tian Yu-Ping, *T-AC Mar 94* 554-558
- Feng Zheng, Mian Chey, and Wei-Bing Gao Feedback stabilization of linear systems with distributed delays in state and control variables, *T-AC Aug 94* 1714-1718
- Fernandes, C., L Gurvits, and Zexiang Li Near-optimal nonholonomic motion planning for a system of coupled rigid bodies, *T-AC Mar 94* 450-463
- Ferrante, A. A parameterization of minimal stochastic realizations, *T-AC Oct 94* 2122-2126
- Ferreira, P.M.G. Comments on "System zeros determination from an unreduced matrix fraction description" (by K S Yeung and C-M Kwan, Nov 93 1695-1697, *T-AC Nov 94* 2367
- Feuer, A. and G C Goodwin Generalized sample hold functions-frequency domain analysis of robustness, sensitivity, and intersample difficulties, *T-AC May 94* 1042-1047
- Feuer, A. see Tuch, J, *T-AC Apr 94* 823-827
- Fialho, I.J., and T T Georgiou On stability and performance of sampled-data systems subject to wordlength constraint, *T-AC Dec 94* 2476-2481
- Fiala, J., and R Lumia The effect of time delay and discrete control on the contact stability of simple position controllers, *T-AC Apr 94* 870-873
- Figueroa, J.L., and J A Romagnoli An algorithm for robust pole assignment via polynomial approach, *T-AC Apr 94* 831-835
- Fiodorov, E.D. Least absolute values estimation computational aspects, *T-AC Mar 94* 626-630
- Fjellstad, O.-E., and T I Fossen Comments on "The attitude control problem", *T-AC Mar 94* 699-700
- Flamm, D.S., and Hong Yang Optimal mixed sensitivity for SISO-distributed plants, *T-AC Jun 94* 1150-1165
- Foo, Y.K., and Y C Soh Closed-loop hyperstability of interval plants, *T-AC Jan 94* 151-154
- Fossen, T.I. see Fjellstad, O.-L *T-AC Mar 94* 699-700
- Fourquet, J.-Y., see Soueres, P, *T-AC Aug 94* 1626-1630
- Frank, P.M. see Ding, X *T-AC Aug 94* 1648-1652
- Franklin, G.F., see Pao L Y, *I-AC Sep 94* 1963-1966
- Freiling, G., see Abou-Kandil, H, *T-AC Aug 94* 1631-1636
- Fu, L.-C. see Lian K -Y, *T-AC Dec 94* 2426-2441
- Fu-Chuang Chen, and Chen-Chung Liu Adaptively controlling nonlinear continuous-time systems using multilayer neural networks, *T-AC Jun 94* 1306-1310

G

- Gajic, Z. see Aganovic, Z, *T-AC Sep 94* 1932-1935
- Gajic, Z. and M T Lim A new filtering method for linear singularly perturbed systems, *T-AC Sep 94* 1952-1955
- Gang Feng. A robust approach to adaptive control algorithms, *T-AC Aug 94* 1738-1742
- Gang Feng, and M Palaniswami Robust direct adaptive controllers with a new normalization technique, *T-AC Nov 94* 2330-2334
- Gang Tao, and P V Kokotovic Adaptive control of plants with unknown dead-zones, *T-AC Jan 94* 59-68
- Gapaki, P.B. and J C Geromel A convex approach to the absolute stability problem, *T-AC Sep 94* 1929-1932
- Garnett, J. S Dasgupta, and C R Johnson, Jr Convergence of the signed output error adaptive identifier, *T-AC Jul 94* 1387-1399
- Garg Huang, see Tsai, W K, *T-AC Mar 94* 534-540
- Ge, Y., E G Collins, Jr, L T Watson, and L D Davis An input normal form homotopy for the L^2 optimal model order reduction problem, *T-AC Jun 94* 1302-1305
- Georgiou, T.T., see Fialho, I J, *T-AC Dec 94* 2476-2481
- Geromel, J.C., see Peres, P L D, *T-AC Jan 94* 198-202
- Geromel, J.C., P L D Peres, and S R Souza H_∞ control of discrete-time uncertain systems, *T-AC May 94* 1072-1075
- Geromel, J.C., see Gapaki, P B, *T-AC Sep 94* 1929-1932
- Gessing, R. Comments about frequency response plots of Keller and Anderson and Rattan, *T-AC Aug 94* 1770-1771
- Ghwan-Lu Tseng, I-Kong Fong, and Juing-Huei Su Analysis and applications of robust nonsingularity problem using the structured singular value, *T-AC Oct 94* 2118-2122
- Gil, M.I. On absolute stability of differential-delay systems, *T-AC Dec 94* 2481-2484
- Gin-Hol Wu, see Wen-June Wang, *T-AC Jan 94* 99-102

- Jia, A., and F. DiCesare. Blocking and controllability of Petri nets in supervisory control; *T-AC Apr 94* 818-823
- Glaria, J.J., and G.C. Goodwin. A parameterization for the class of all stabilizing controllers for linear minimum phase plants; *T-AC Feb 94* 433-434
- Glover, K., see Kemin Zhou, *T-AC Aug 94* 1564-1574
- Glover, K., see Doyle, J., *T-AC Aug 94* 1575-1587
- Glumineau, A., see Di Benedetto, M.D., *T-AC Jun 94* 1246-1250
- Godhavn, J.-M., see Egeland, O., *T-AC Apr 94* 842-846
- Godhavn, J.-M., see Egeland, O., *T-AC Aug 94* 1671-1673
- Golovan, A., and A. Matasov. The Kalman-Bucy filter in the guaranteed estimation problem; *T-AC Jun 94* 1282-1286
- Gong, C., and S. Thompson. Stability margin evaluation for uncertain linear systems; *T-AC Mar 94* 548-550
- Gonzalez, G.A., M.T. Tzou, and C.E. D'Attellis. A remark on chaotic behavior in adaptive control systems; *T-AC Oct 94* 2145-2148
- Goodwin, G.C., see Glaria, J.J., *T-AC Feb 94* 433-434
- Goodwin, G.C., see Turan, L., *T-AC Mar 94* 601-605
- Goodwin, G.C., see Feuer, A., *T-AC May 94* 1042-1047
- Goodwin, G.C., see Weller, S.R., *T-AC Jul 94* 1360-1375
- Gorez, R., see Bandyopadhyay, B., *T-AC Dec 94* 2454-2456
- Grabbe, M.T., and M.M. Bridges. Comments on "Force/motion control of constrained robots using sliding mode"; *T-AC Jan 94* 179
- Greblicki, W. Nonparametric identification of Wiener systems by orthogonal series; *T-AC Oct 94* 2077-2086
- Grimble, M.J. Two and a half degrees of freedom LQG controller and application to wind turbines; *T-AC Jan 94* 122-127
- Grizzle, J.W. Review of "Nonlinear Systems" (Khalil, H.; 1992); *T-AC Jan 94* 251-252
- Grizzle, J.W., see Di Benedetto, M.D., *T-AC Aug 94* 1539-1550
- Grizzle, J.W., M.D. Di Benedetto, and F. Lamnabhi-Lagarigue. Necessary conditions for asymptotic tracking in nonlinear systems; *T-AC Sep 94* 1782-1794
- Gu, K. H_∞ control of systems under norm bounded uncertainties in all system matrices; *T-AC Jun 94* 1320-1322
- Guanghan Xu, see Young Man Cho, *T-AC Oct 94* 2004-2017
- Guang-Ren Duan. Eigenstructure assignment by decentralized output feedback: a complete parametric approach; *T-AC May 94* 1009-1014
- Gudmundsson, T., and A.J. Laub. Approximate solution of large sparse Lyapunov equations; *T-AC May 94* 1110-1114
- Guillaume, P., see Pintelon, R., *T-AC Nov 94* 2245-2260
- Gundes, A.N. Stabilizing controller design for linear systems with sensor or actuator failures; *T-AC Jun 94* 1224-1230
- Guo, L., see Ding, X., *T-AC Aug 94* 1648-1652
- Guoxiang Gu. Suboptimal algorithms for worst case identification in H^∞ and model validation; *T-AC Aug 94* 1657-1661
- Guo-Xiang Yu, see Jian-Qiang Hu, *T-AC Sep 94* 1875-1880
- Gurvits, L., see Fernandes, C., *T-AC Mar 94* 450-463
- Gutman, P.-O., C. Baril, and L. Neumann. An algorithm for computing value sets of uncertain transfer functions in factored real form; *T-AC Jun 94* 1268-1273

H

- Habib, N.R., see Premaratne, K., *T-AC Mar 94* 581-585
- Haddad, W.M., Hsing-Hsin Huang, and D.S. Bernstein. Sampled-data observers with generalized holds for unstable plants; *T-AC Jan 94* 229-234
- Haddad, W.M., D.S. Bernstein, and Y.W. Wang. Dissipative H_2/H_∞ controller synthesis; *T-AC Apr 94* 827-831
- Haddad, W.M., see Bernstein, D.S., *T-AC Jul 94* 1513-1517
- Haddad, W.M., and D.S. Bernstein. Corrections to "Robust stability and performance via fixed-order dynamic compensation: The discrete-time case" (May 93 776-782); *T-AC Aug 94* 1772
- Haddad, W.M., and R. Moser. Optimal dynamic output feedback for nonzero set point regulation: the discrete-time case; *T-AC Sep 94* 1921-1925
- Haddad, W.M., see Wang, Y.W., *T-AC Nov 94* 2284-2287
- Hadj-Alouane, N.B., S. Lafortune, and F. Lin. Variable lookahead supervisory control with state information; *T-AC Dec 94* 2398-2410
- Hajeer, H.Y., see Mahmoud, M.S., *T-AC Jan 94* 148-151
- Halevi, Y. Stable LQG controllers; *T-AC Oct 94* 2104-2106
- Hall, S.R. Comments on two methods for designing a digital equivalent to a continuous control system; *T-AC Feb 94* 420-421
- Halpern, M.E. Preview tracking for discrete-time SISO systems; *T-AC Mar 94* 589-592
- Hammer, J. Internally stable nonlinear systems with disturbances: a parameterization; *T-AC Feb 94* 300-314
- Han-Fu Chen, and Qian-Yu Tang. Stability analysis for manufacturing systems with unreliable machines and random inputs; *T-AC Mar 94* 681-686
- Hang, C.C., see Astrom, K.J., *T-AC Feb 94* 343-345

- Han Ho Choi, and Myung Jin Chung. Estimation of the asymptotic stability region of uncertain systems with bounded sliding mode controllers; *T-AC Nov 94* 2275-2278
- Hara, S., see Hayakawa, Y., *T-AC Nov 94* 2278-2284
- Hayakawa, Y., S. Hara, and Y. Yamamoto. H_∞ type problem for sampled-data control systems—a solution via minimum energy characterization; *T-AC Nov 94* 2278-2284
- Hayton, G.E., see Pugh, A.C., *T-AC May 94* 1141-1145
- Heiss, M. Inverse passive learning of an input-output map through update-spline-smoothing; *T-AC Feb 94* 259-268
- Helferty, J.J., see Jakubowski, A.M., *T-AC May 94* 1145-1147
- Hellal, A., see Aggoune, M.E., *T-AC Oct 94* 2106-2110
- Hench, J.J., and A.J. Laub. Numerical solution of the discrete-time periodic Riccati equation; *T-AC Jun 94* 1197-1210
- Hendel, B., see Limebeer, D.J.N., *T-AC Jan 94* 69-82
- Herrera, A., see Ortega, R., *T-AC Aug 94* 1639-1643
- Hmamed, A. Comments, with reply, on "Vector norms as Lyapunov functions for linear systems" (by H. Kiendl et al., Jun 92 839-842); *T-AC Dec 94* 2522-2523
- Hoai Nghia Duong, and I.D. Landau. On test horizon for model validation by output error; *T-AC Jan 94* 102-106
- Hoai Nghia Duong, and I.D. Landau. On statistical properties of a test for model structure selection using the extended instrumental variable approach; *T-AC Jan 94* 211-215
- Hodel, A.S., see Davis, L.D., *T-AC Apr 94* 849-852
- Holiot, C.V., and R. Tempo. On the Nyquist envelope of an interval plant family; *T-AC Feb 94* 391-396
- Holmberg, U., see Myszkowski, P., *T-AC Nov 94* 2366-2367
- Hong Jiang, see Lin-Zhang Lu, *T-AC Aug 94* 1682-1685
- Hong Yang, see Flamm, D.S., *T-AC Jun 94* 1150-1165
- Hooker, M.A. Analytic first and second derivatives for the recursive prediction error algorithm's log likelihood function; *T-AC Mar 94* 662-664
- Hopkins, W.E., Jr. Exponential linear quadratic optimal control with discounting; *T-AC Jan 94* 175-178
- Hong-Glou Chen, and Kuang-Wei Han. Improved quantitative measures of robustness for multivariable systems; *T-AC Apr 94* 807-810
- Hou, M., and P.C. Muller. Disturbance decoupled observer design: a unified viewpoint; *T-AC Jun 94* 1338-1341
- Hsi-Han Yeh, see Hsu, C.S., *T-AC Aug 94* 1679-1681
- Hsing-Hsin Huang, see Haddad, W.M., *T-AC Jan 94* 229-234
- Hsu, C.S., Xianggang Yu, Hsi-Han Yeh, and S.S. Banda. H_∞ compensator design with minimal order observers; *T-AC Aug 94* 1679-1681
- Huang, G.M., see Antonio, J.K., *T-AC Feb 94* 385-391
- Huang, G.M., see Antonio, J.K., *T-AC Sep 94* 1839-1844
- Hua Xu, and K. Mizukami. Linear-quadratic zero-sum differential games for generalized state space systems; *T-AC Jan 94* 143-147
- Hua Xu, and K. Mizukami. New sufficient conditions for linear feedback closed-loop Stackelberg strategy of descriptor systems; *T-AC May 94* 1097-1102
- Humes, C., Jr. A regulator stabilization technique: Kumar-Seidman revisited; *T-AC Jan 94* 191-196
- Hunt, K.J., M. Sebek, and V. Kucera. Polynomial solution of the standard multivariable H_2 -optimal control problem; *T-AC Jul 94* 1502-1507
- Hunt, L.R., and M.S. Verma. Observers for nonlinear systems in steady state; *T-AC Oct 94* 2113-2118

I

- I-Kong Fong, see Juing-Huei Su, *T-AC Jun 94* 1341-1344
- I-Kong Fong, see Ghwan-Lu Tseng, *T-AC Oct 94* 2118-2122
- Imura, J.-I., T. Sugie, and T. Yoshikawa. Global robust stabilization of nonlinear cascaded systems; *T-AC May 94* 1084-1089
- In-Joong Ha, and Chang-Ik Kang. Explicit characterization of all feedback-linearizing controllers for a general type brushless DC motor; *T-AC Mar 94* 673-677
- In-Joong Ha, and Sung-Joon Lee. Input-output linearization with state equivalence and decoupling; *T-AC Nov 94* 2269-2274
- Ioannou, P. Corrections to "On the stability proof of adaptive schemes with static normalizing signal and parameter projection" (Jan 93 170-173); *T-AC Apr 94* 896
- Ioannou, P.A., see Dattam, A., *T-AC Dec 94* 2370-2387
- Irving, W.W., and J.N. Tsitsiklis. Some properties of optimal threshold decentralized detection; *T-AC Apr 94* 835-838
- Ishijima, S., see Kojima, A., *T-AC Aug 94* 1694-1698
- Ismail, O., see Bandyopadhyay, B., *T-AC Dec 94* 2454-2456
- Ito, S., see Shimizu, K., *T-AC May 94* 982-986

J

- Jaecheong Park, and G. Rizzoni. An eigenstructure assignment algorithm for the design of fault detection filters; *T-AC Jul 94* 1521-1524
- Jaensch, C.R., see Chih-Hai Fan, *T-AC Oct 94* 1986-2003
- Jakubowski, A.M., and J.J. Heflerty. Comments on "Block multirate input-output model for sampled-data control systems"; *T-AC May 94* 1145-1147
- James, M.R., J.S. Baras, and R.J. Elliott. Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems; *T-AC Apr 94* 780-792
- James, M.R. On the certainty equivalence principle and the optimal control of partially observed dynamic games; *T-AC Nov 94* 2321-2324
- Jang Gyu Lee, see Dohyoung Chung, *T-AC Aug 94* 1698-1702
- Janiak, A., see Cheng, T.C.E., *T-AC Jun 94* 1243-1246
- Jank, G., see Abou-Kandil, H., *T-AC Aug 94* 1631-1636
- Jaulin, L., see Walter, E., *T-AC Apr 94* 886-889
- Jetto, L. Deadbeat ripple-free tracking with disturbance rejection: A polynomial approach; *T-AC Aug 94* 1759-1764
- Jiang, J., and S.X.C. Lou. Production control of manufacturing systems: A multiple time scale approach; *T-AC Nov 94* 2292-2297
- Jian-Qiang Hu, and Dong Xiang. Structural properties of optimal production controllers in failure-prone manufacturing systems; *T-AC Mar 94* 640-643
- Jian-Qiang Hu, P. Vakili, and Guo-Xiang Yu. Optimality of hedging point policies in the production control of failure prone manufacturing systems; *T-AC Sep 94* 1875-1880
- Jie Chen, and H.L. Weinert. Stationarity and reciprocity in stochastic multipoint boundary value systems; *T-AC May 94* 1114-1116
- Jie Huang, and Ching-Fang Lin. On a robust nonlinear servomechanism problem; *T-AC Jul 94* 1510-1513
- Jie Huang, and Ching-Fang Lin. A stability property and its application to discrete-time nonlinear system control; *T-AC Nov 94* 2307-2311
- Ji Feng Zhang, see Xi Sun, *T-AC Jan 94* 207-211
- Ji Feng Zhang. Comments on "Robust adaptive regulation with minimal prior knowledge"; *T-AC Mar 94* 605
- Jin-Hoon Kim, and Zeungnam Bien. Robust stability of uncertain linear systems with saturating actuators; *T-AC Jan 94* 202-207
- Jinn-Wen Wu, see Keum-Shik Hong, *T-AC Jul 94* 1432-1436
- Johnson, C.R., Jr., see Garnett, J., *T-AC Jul 94* 1387-1399
- Jong-Hwan Kim, see Keun-Mo Koo, *T-AC Jun 94* 1230-1233
- Joon Hwa Lee, Sang Woo Kim, and Wook Hyun Kwon. Memoryless H^∞ controllers for state delayed systems; *T-AC Jan 94* 159-162
- Juang, J.N., see Morris, K.A., *T-AC May 94* 1056-1063
- Juditsky, A., and P. Priouret. A robust algorithm for random parameter tracking; *T-AC Jun 94* 1211-1221
- Juing-Huei Su, I-Kong Fong, and Chwan-Lu Tseng. Stability analysis of linear systems with time delay; *T-AC Jun 94* 1341-1344
- Juing-Huei Su, see Ghwan-Lu Tseng, *T-AC Oct 94* 2118-2122
- Ju-Jang Lee, and Yangsheng Xu. A new method of switching surface design for multivariable variable structure systems; *T-AC Feb 94* 414-419
- Jury, E.I., see Premaratne, K., *T-AC Feb 94* 352-355

K

- Kabamba, P.T., see Shuoh Rern, *T-AC Jan 94* 162-164
- Kabamba, P.T., see Chang Yang, *T-AC Mar 94* 657-661
- Kabamba, P.T., S.M. Meerkov, and E.-K. Poh. H_2 -optimal zeros; *T-AC Jun 94* 1298-1301
- Kabamba, P.T., S.M. Meerkov, and E.-K. Poh. Youla parameterization in closed-loop vibrational control; *T-AC Jul 94* 1455-1459
- Kailath, T., see Pal, D., *T-AC Jan 94* 238-245
- Kailath, T., see Sayed, A.H., *T-AC Mar 94* 619-623
- Kailath, T., see Sayed, A.H., *T-AC May 94* 960-976
- Kailath, T., see Constantinescu, T., *T-AC Jul 94* 1476-1481
- Kailath, T., see Young Man Cho, *T-AC Oct 94* 2004-2017
- Kailath, T., see Sayed, A.H., *T-AC Nov 94* 2265-2269
- Kaining Wang, A.N. Michel, and Derong Liu. Necessary and sufficient conditions for the Hurwitz and Schur stability of interval matrices; *T-AC Jun 94* 1251-1255
- Kaining Wang, and A.N. Michel. Necessary and sufficient conditions for the controllability and observability of a class of linear, time-invariant systems with interval plants; *T-AC Jul 94* 1443-1447
- Kouptelidis, N., and Tsinias, J., *T-AC Apr 94* 806
- Krener, E.W. Review of "Applied Optimal Control and Estimation" (Lewis, F.L.; 1992); *T-AC Aug 94* 1773-1774
- Krasovskiy, R.D., and T.E. Djaferis. A novel approach to the analysis and synthesis of controllers for parametrically uncertain systems; *T-AC Apr 94* 874-876
- Kraoun, M.B.A., see Derbel, N., *T-AC Jul 94* 1429-1431
- Kanellakopoulos, I., see Krstic, M., *T-AC Apr 94* 738-752
- Kanellakopoulos, I., see Marino, R., *T-AC Jun 94* 1314-1319

- Kanellakopoulos, I., and D.M. Wiberg. Review of "Adaptive Control Systems" (Isermann, R., et al.; 1992); *T-AC Aug 94* 1776
- Kanellakopoulos, I. A discrete-time adaptive nonlinear system; *T-AC Nov 94* 2362-2365
- Kanestrom, R.K., and O. Egeland. Nonlinear active vibration damping; *T-AC Sep 94* 1925-1928
- Karampetakis, N.P., see Pugh, A.C., *T-AC May 94* 1141-1145
- Karan, M., M.E. Sezer, and O. Ocali. Robust stability of discrete-time systems under parametric perturbations; *T-AC May 94* 991-995
- Karason, S.P., and A.M. Annaswamy. Adaptive control in the presence of input constraints; *T-AC Nov 94* 2325-2330
- Karcanias, N., and M. Mitrouli. A matrix pencil based numerical method for the computation of the GCD of polynomials; *T-AC May 94* 977-981
- Karl, W.C., G.C. Verghese, and J.H. Lang. Control of vibrational systems; *T-AC Jan 94* 222-226
- Karray, F., V.J. Modi, and T.A.W. Dwyer. On the elastic mode estimation aspect of a class of multibody flexible systems; *T-AC May 94* 1016-1020
- Kataoka, S., see Yamada, T., *T-AC Mar 94* 696-698
- Kavranoglu, D. H_∞ -norm approximation of systems by constant matrices and related results; *T-AC May 94* 1006-1009
- Kavranoglu, D. A computational scheme for H_∞ model reduction; *T-AC Jul 94* 1447-1451
- Kavranoglu, D., and M. Bettayeb. Characterization and computation of the solution to the optimal L_∞ approximation problem; *T-AC Sep 94* 1899-1904
- Kazakos, D., and J. Tsinias. The input-to-state stability condition and global stabilization of discrete-time systems; *T-AC Oct 94* 2111-2113
- Keel, L.H., and S.P. Bhattacharyya. Robust parametric classical control design; *T-AC Jul 94* 1524-1530
- Kehui Wei. Stabilization of linear time-invariant interval systems via constant state feedback control; *T-AC Jan 94* 22-32
- Kelly, R., R. Ortega, A. Ailon, and A. Loria. Global regulation of flexible joint robots using approximate differentiation; *T-AC Jun 94* 1222-1224
- Kemin Zhou, K. Glover, B. Bodenheimer, and J. Doyle. Mixed H_2 and H_∞ performance objectives. I. Robust performance analysis; *T-AC Aug 94* 1564-1574
- Keqin Gu. Designing stabilizing control of uncertain systems by quasicontex optimization; *T-AC Jan 94* 127-131
- Keqin Gu. Comments on "convexity of frequency response arcs associated with a stable polynomial"; *T-AC Nov 94* 2262-2265
- Keum-Shik Hong, see Wu, J.W., *T-AC Apr 94* 811-814
- Keum-Shik Hong, Jinn-Wen Wu, and Kyo-II Lee. New conditions for the exponential stability of evolution equations; *T-AC Jul 94* 1432-1436
- Keum Shik Hong, and J. Bentsman. Direct adaptive control of parabolic systems. algorithm synthesis and convergence and stability analysis; *T-AC Oct 94* 2018-2033
- Keun-Mo Koo, and Jong-Hwan Kim. Robust control of robot manipulators with parametric uncertainty; *T-AC Jun 94* 1230-1233
- Khalil, H.K. Review of "Linear System Theory" (Rugh, W.J.; 1993); *T-AC Dec 94* 2528-2529
- Khargonekar, P., see Poola, K., *T-AC May 94* 951-959
- Khargonekar, P.P., and A.B. Ozguler. Decentralized control and periodic feedback; *T-AC Apr 94* 877-882
- Khargonekar, P.P., see Weigand Sun, *T-AC Oct 94* 2034-2046
- Kharitonov, V.L., and A.P. Zhabko. Robust stability of time-delay systems; *T-AC Dec 94* 2388-2397
- Khayatian, A., and D.G. Taylor. Feedback control of linear systems by multirate pulse-width modulation; *T-AC Jun 94* 1292-1297
- Khayatian, A., and D.G. Taylor. Multirate modeling and control design for switched-mode power converters; *T-AC Sep 94* 1848-1852
- Khorasani, K. A robust adaptive control design for a class of dynamical systems using corrected models; *T-AC Aug 94* 1726-1732
- Khosla, P., see Bodson, M., *T-AC Sep 94* 1939-1944
- Kleinman, D.L., see Pete, A., *T-AC Aug 94* 1702-1707
- Koc, C.K., B. Bakkaloglu, and L.S. Shieh. Computation of the matrix sign function using continued fraction expansion; *T-AC Aug 94* 1644-1647
- Kogan, J. Robust performance of systems with affine parameter uncertainty and convex analysis; *T-AC Jan 94* 227-229
- Koh, E.K., see Lee, T.H., *T-AC Mar 94* 565-568
- Kojima, A., K. Uchida, E. Shimemura, and S. Ishijima. Robust stabilization of a system with delays in control; *T-AC Aug 94* 1694-1698
- Kokotovic, P.V., see Gang Tao, *T-AC Jan 94* 59-68
- Kokotovic, P.V., see Krstic, M., *T-AC Apr 94* 738-752
- Kokotovic, P.V., see Marino, R., *T-AC Jun 94* 1314-1319
- Komaroff, N. Diverse bounds for the eigenvalues of the continuous algebraic Riccati equation; *T-AC Mar 94* 532-534
- Komaroff, N. Iterative matrix bounds and computational solutions to the discrete algebraic Riccati equation; *T-AC Aug 94* 1676-1678
- Koent, R.L., see Massoumnia, M.-A., *T-AC Aug 94* 1027-1031
- Koumboulis, F.N., see Paraskevopoulos, P.N., *T-AC Jan 94* 185-190
- Kozin, F., see Zhi Yu Zhang, *T-AC Mar 94* 560-565

- Krajewski, W., A. Lepschy, and U. Viaro. Reduction of linear continuous-time multivariable systems by matching first- and second-order information; *T-AC Oct 94* 2126-2129
- Krause, J., see Poolla, K., *T-AC May 94* 951-959
- Kreisselmeier, G., and T. Birkholzer. Numerical nonlinear regulator design; *T-AC Jan 94* 33-46
- Kreisselmeier, G. Parameter adaptive control: a solution to the overmodeling problem; *T-AC Sep 94* 1819-1826
- Krishnamurthy, V. On-line estimation of dynamic shock-error models based on the Kullback-Leibler information measure; *T-AC May 94* 1129-1135
- Krishnan, H., and N.H. McClamroch. On the connection between nonlinear differential-algebraic equations and singularly perturbed control systems in nonstandard form; *T-AC May 94* 1079-1084
- Krstic, M., I. Kanellakopoulos, and P.V. Kokotovic. Nonlinear design of adaptive controllers for linear systems; *T-AC Apr 94* 738-752
- Kuang-Wei Han, see Horng-Giou Chen, *T-AC Apr 94* 807-810
- Kucera, V., see Hunt, K.J., *T-AC Jul 94* 1502-1507
- Kumar, P.R., see Ellis, R.D., *T-AC Sep 94* 1795-1801
- Kumar, P.R., see Wei Ren, *T-AC Oct 94* 2047-2060
- Kwakernaak, H., and M. Sebek. Polynomial J-spectral factorization; *T-AC Feb 94* 315-328
- Kwanghee Nam, Seongno Lee, and Sangchul Won. A local stabilizing control scheme using an approximate feedback linearization; *T-AC Nov 94* 2311-2314
- Kyo-Il Lee, see Keum-Shik Hong, *T-AC Jul 94* 1432-1436

L

- Lafortune, S., see Hadj-Alouane, N.B., *T-AC Dec 94* 2398-2410
- Lai, H.Y., see Song, Y.D., *T-AC Sep 94* 1866-1871
- Lam, J. Analysis on the Laguerre formula for approximating delay systems; *T-AC Jul 94* 1517-1521
- Lamnabhi-Lagarigue, F., see Grizzle, J.W., *T-AC Sep 94* 1782-1794
- Landau, I.D., see Hoai Nghia Duong, *T-AC Jan 94* 102-106
- Landau, I.D., see Hoai Nghia Duong, *T-AC Jan 94* 211-215
- Landi, A., see Balestrino, A., *T-AC Jun 94* 1255-1258
- Lang, J.H., see Karl, W.C., *T-AC Jan 94* 222-226
- Lang Hong. Multiresolutional distributed filtering; *T-AC Apr 94* 853-856
- Lang Zi-Qiang. On identification of the controlled plants described by the Hammerstein system; *T-AC Mar 94* 569-573
- Laub, A.J., see Gudmundsson, T., *T-AC May 94* 1110-1114
- Laub, A.J., see Hench, J.J., *T-AC Jun 94* 1197-1210
- Laub, A.J., see Erickson, M.A., *T-AC Sep 94* 1904-1909
- Laumond, J.P., see Walsh, G., *T-AC Jan 94* 216-222
- Laumond, J.-P., see Soueres, P., *T-AC Aug 94* 1626-1630
- Lawrence, D.A., and W.J. Rugh. Input-output pseudolinearization for nonlinear systems; *T-AC Nov 94* 2207-2218
- Lawrence, D.A., see Radenkovic, M.S., *T-AC Nov 94* 2357-2361
- Lee, T.H., Q.G. Wang, and E.K. Koh. An iterative algorithm for pole placement by output feedback; *T-AC Mar 94* 565-568
- LeGall, J.P., see Premaratne, K., *T-AC Mar 94* 581-585
- Lehman, B., J. Bentsman, S.V. Lunel, and E.I. Verriest. Vibrational control of nonlinear time lag systems with bounded delay: averaging theory, stabilizability, and transient behavior; *T-AC May 94* 898-912
- Lehman, B., and K. Shujee. Delay independent stability conditions and decay estimates for time-varying functional differential equations; *T-AC Aug 94* 1673-1676
- Leland, R.P. Stability of asynchronous systems with Poisson transitions; *T-AC Jan 94* 182-185
- Leon de la Barra S., B.A. Frequency domain tradeoffs in loop transfer recovery for multivariable nonminimum phase discrete-time systems; *T-AC Mar 94* 574-577
- Leon de la Barra S., B.A. On undershoot in SISO systems; *T-AC Mar 94* 578-581
- Lepschy, A., see Krajewski, W., *T-AC Oct 94* 2126-2129
- Lepschy, A., see Beghi, A., *T-AC Dec 94* 2494-2496
- Lev-Ari, H., see Sayed, A.H., *T-AC Nov 94* 2265-2269
- Levin, A. An analytical method of estimating the domain of attraction for polynomial differential equations; *T-AC Dec 94* 2471-2475
- Lewis, F.L., see Symos, V.L., *T-AC Feb 94* 410-414
- Lewis, F.L. Review of "Linear Multivariable Control: Algebraic Analysis and Synthesis Methods" (Vardulakis, A.J.G.; 1991); *T-AC Jul 94* 1536
- Le Yi Wang, see Zames, G., *T-AC Sep 94* 1827-1838
- Li, Y., and W.M. Wonham. Correction to "Control of vector discrete-event systems I - The base model" (Aug 93 1214-1227); *T-AC Aug 94* 1771
- Lian, K.-Y., L.-S. Wang, and L.-C. Fu. Controllability of spacecraft systems in a central gravitational field; *T-AC Dec 94* 2426-2441
- Libropoulos, G., and M. Caramanis. Production control of manufacturing systems with production rate-dependent failure rates; *T-AC Apr 94* 889-895
- Lihua Xie, Yeng Chai Soh, and C.E. de Souza. Robust Kalman filtering for uncertain discrete-time systems; *T-AC Jun 94* 1310-1314

- Li Lee, see Chun-Hsiung Fang, *T-AC Nov 94* 2303-2307
- Lim, B.C., see Astrom, K.J., *T-AC Feb 94* 343-345
- Lim, M.T., see Gajic, Z., *T-AC Sep 94* 1952-1955
- Limebeer, D.J.N., B.D.O. Anderson, and B. Hendel. A Nash game approach to mixed H_2/H_∞ control; *T-AC Jan 94* 69-82
- Lin, F., and H. Mortazavian. A normality theorem for decentralized control of discrete-event systems; *T-AC May 94* 1089-1093
- Lin, F., see Hadj-Alouane, N.B., *T-AC Dec 94* 2398-2410
- Lin, L., see Zames, G., *T-AC Sep 94* 1827-1838
- Lin-Zhang Lu, Xingzhi Ji, and Hong Jiang. An algorithm using the HR process for solving the closed-loop eigenvalues of a discrete-time algebraic Riccati equation; *T-AC Aug 94* 1682-1685
- Liu Danyang, and Liu Xuanhuang. Optimal state estimation without the requirement of a priori statistics information of the initial state; *T-AC Oct 94* 2087-2091
- Liu Hsu, A.D. de Araujo, and R.R. Costa. Analysis and design of I/O based variable structure adaptive control; *T-AC Jan 94* 4-21
- Liu Xuanhuang, see Liu Danyang, *T-AC Oct 94* 2087-2091
- Livstone, M.M., J.A. Farrell, and M.A. Dahleh. Comments on "Least squares methods for H_∞ control oriented system identification"; *T-AC Jul 94* 1531
- Li-Wen Chen, and G.P. Papavassilopoulos. Robust variable structure and switching- Σ adaptive control of single-arm dynamics; *T-AC Aug 94* 1621-1626
- Li-Xiu Wang. A supervisory controller for fuzzy control systems that guarantees stability; *T-AC Sep 94* 1845-1847
- Logemann, H., and L. Pandolfi. A note on stability and stabilizability of neutral systems; *T-AC Jan 94* 138-143
- Longhi, S. Structural properties of multirate sampled-data systems; *T-AC Mar 94* 692-696
- Loparo, K.A., see Fang, Y., *T-AC Dec 94* 2489-2490
- Loria, A., see Kelly, R., *T-AC Jun 94* 1222-1224
- Lou, S.X.C., see Jiang, J., *T-AC Nov 94* 2292-2297
- Lovass-Nagy, V., D.L. Powers, and R.J. Schilling. On regularizing descriptor systems by output feedback; *T-AC Jul 94* 1507-1509
- Lozano, R., and Xiao-Hui Zhao. Adaptive pole placement without excitation probing signals; *T-AC Jan 94* 47-58
- Lozano, R., A. Osorio, and J. Torres. Adaptive stabilization of nonminimum phase first-order continuous-time systems; *T-AC Aug 94* 1748-1751
- Lozano, R., see Brogliato, B., *T-AC Aug 94* 1764-1768
- Lozano, R., see Moctezuma, R.G., *T-AC Sep 94* 1856-1860
- Lu, W.-M., and J.C. Doyle. H_∞ control of nonlinear systems via output feedback. Controller parameterization; *T-AC Dec 94* 2517-2524
- Lucibello, P. A note on a necessary condition for output regulation; *T-AC Mar 94* 558-559
- Lucibello, P. Repositioning control of robotic arms by learning; *T-AC Aug 94* 1690-1694
- Lumia, R., see Fiala, J., *T-AC Apr 94* 870-873
- Lunel, S.V., see Lehman, B., *T-AC May 94* 898-912
- Luzeaux, D. Process control and machine learning rule-based incremental control; *T-AC Jun 94* 1166-1171

M

- Magni, J.-F., and P. Mouyon. On residual generation by observer and parity space approaches; *T-AC Feb 94* 441-447
- Mahmoud, M.S., and H.Y. Hajeer. A globally convergent adaptive controller for robot manipulators; *T-AC Jan 94* 148-151
- Mahmoud, M.S., and N.F. Al-Muthairi. Design of robust controllers for time-delay systems; *T-AC May 94* 995-999
- Mahmoud, M.S., and N.F. Al-Muthairi. Quadratic stabilization of continuous time systems with state-delay and norm-bounded time-varying uncertainties; *T-AC Oct 94* 2135-2139
- Mahmoud, M.S. Stabilizing control for a class of uncertain interconnected systems; *T-AC Dec 94* 2484-2488
- Makila, P.M., see Partington, J.R., *T-AC Oct 94* 2171-2176
- Makowski, A.M., see Dye-Jyun Ma, *T-AC Jul 94* 1439-1442
- Malabre, M., see Martinez Garcia, J.C., *T-AC Dec 94* 2457-2460
- Malhame, R.P., see El-Ferik, S., *T-AC Jun 94* 1184-1196
- Mandl, P., see Duncan, T.E., *T-AC Jan 94* 118-122
- Mansour, M., see Sreeram, V., *T-AC Feb 94* 381-385
- Man Zhihong, and M. Palaniswami. Robust tracking control for rigid robot manipulators; *T-AC Jan 94* 154-159
- Mareels, I.M.Y., see Weyer, E., *T-AC Aug 94* 1665-1671
- Marino, R., P. Tomei, I. Kanellakopoulos, and P.V. Kokotovic. Adaptive tracking for a class of feedback linearizable systems; *T-AC Jun 94* 1314-1319
- Marquez, H.J., and C.P. Diduch. Absolute stability of systems with parametric uncertainty and nonlinear feedback; *T-AC Mar 94* 664-668
- Martinez Garcia, J.C., and M. Malabre. The row by row decoupling problem with stability: A structural approach; *T-AC Dec 94* 2457-2460

- Massoumnia, M.-A., and R.L. Kosut A family of norms for system identification problems, *T-AC May 94* 1027-1031
- Matasov, A., see Golovan, A., *T-AC Jun 94* 1282-1286
- Matasov, A.I. The Kalman-Bucy filter accuracy in the guaranteed parameter estimation problem with uncertain statistics, *T-AC Mar 94* 635-639
- McClamroch, N.H., see Krishnan, H., *T-AC May 94* 1079-1084
- McFarlane, D.C., see Petersen, I.R., *T-AC Sep 94* 1971-1977
- Meerkov, S.M., see Kabamba, P.T., *T-AC Jun 94* 1298-1301
- Meerkov, S.M., see Kabamba, P.T., *T-AC Jul 94* 1455-1459
- Megretski, A., see Rantzer, A., *T-AC Sep 94* 1802-1808
- Mehrmann, V., see Bunse-Gerstner, A., *T-AC Aug 94* 1742-1748
- Mian Chey, see Feng Zheng, *T-AC Aug 94* 1714-1718
- Michel, A.N., see Passino, K.M., *T-AC Feb 94* 269-279
- Michel, A.N., see Kaining Wang, *T-AC Jun 94* 1251-1255
- Michel, A.N., see Kaining Wang, *T-AC Jul 94* 1443-1447
- Michel, A.N., see Passino, K.M., *T-AC Jul 94* 1531
- Miller, D.E. Adaptive stabilization using a nonlinear time-varying controller, *T-AC Jul 94* 1347-1359
- Mingori, D.L., see Turan, L., *T-AC Mar 94* 601-605
- Ming-Tzu Ho, see Datta, A., *T-AC Sep 94* 1977-1980
- Mitchell, T.L., see Song, Y.D., *T-AC Sep 94* 1866-1871
- Mitrouli, M., see Karcianias, N., *T-AC May 94* 977-981
- Miyamoto, S., and S. Arinaga Relationship between availability of states and pole/zero cancellations in H_∞ -control, *T-AC Feb 94* 379-381
- Mizukami, K., see Hua Xu, *T-AC Jan 94* 143-147
- Mizukami, K., see Hua Xu, *T-AC May 94* 1097-1102
- Moctezuma, R.G., and R. Lozano Singularity-free multivariable model reference adaptive control, *T-AC Sep 94* 1856-1860
- Modi, V.J., see Karray, F., *T-AC May 94* 1016-1020
- Moog, C.H., see Di Benedetto, M.D., *T-AC Jun 94* 1246-1250
- Moog, C.H., see Castro-Linares, R., *T-AC Oct 94* 2154-2158
- Mooi Choo Chuah. Analysis of networks of queues via projection techniques, *T-AC Aug 94* 1588-1599
- Moore, E., see Campbell, S.L., *T-AC May 94* 1047-1052
- Moore, K.L. Review of "Feedback Control Theory" (Doyle J. et al., 1992), *T-AC Jul 94* 1532-1534
- Morari, M., see Campo, P.J., *T-AC May 94* 932-944
- Morari, M., see Braatz, R.P., *T-AC May 94* 1000-1002
- Morgul, O., Bo Peng Rao, and F. Conrad On the stabilization of a cable with a tip mass, *T-AC Oct 94* 2140-2145
- Morris, K.A., and J.N. Juang Dissipative controller designs for second-order dynamic systems, *T-AC May 94* 1056-1063
- Morris, K.A. Convergence of controllers designed using state-space techniques, *T-AC Oct 94* 2100-2104
- Morse, A.S., see Pait, F.M., *T-AC Jun 94* 1172-1183
- Morse, A.S., and F.M. Pait MIMO design models and internal regulators for cyclicly switched parameter-adaptive control systems, *T-AC Sep 94* 1809-1818
- Mortazavian, H., see Lin, F., *T-AC May 94* 1089-1093
- Mosca, E., see Chisci, I., *T-AC Mar 94* 614-618
- Mosca, E., see Clarke, D.W., *T-AC May 94* 1052-1056
- Moser, R., see Haddad, W.M., *T-AC Sep 94* 1921-1925
- Mouyon, P., see Magni, J.-F., *T-AC Feb 94* 441-447
- M'Saad, M., see Tadjine, M., *T-AC Jun 94* 1259-1262
- Mukhopadhyay, S., see Bose, S., *T-AC Apr 94* 803-806
- Muller, P.C., see Hou, M., *T-AC Jun 94* 1338-1341
- Murray, R., see Walsh, G., *T-AC Jan 94* 216-222
- Musicki, D., R. Evans, and S. Stankovic Integrated probabilistic data association, *T-AC Jun 94* 1237-1241
- Mutoh, Y., and P.N. Nikiforuk Suboptimal perfect model matching for a plant with measurement noise and its application to MRACS, *T-AC Feb 94* 422-425
- Myszka, W., see Rafajlowicz, E., *T-AC Jun 94* 1241-1243
- Myszkowski, P., and U. Holmberg Comments on "Variable structure control design for uncertain discrete-time systems", *T-AC Nov 94* 2366-2367
- Nung Jin Chung, see Han Ho Choi, *T-AC Nov 94* 2275-2278
- Nichola, N.K., see Bunse-Gerstner, A., *T-AC Aug 94* 1742-1748
- Nie-Zen Yen, see Yung-Chun Wu, *T-AC Jan 94* 112-117
- Nie-Zen Yen, see Yung-Chun Wu, *T-AC Jun 94* 1287-1291
- Nijmeijer, H., see Ruiz, A.C., *T-AC Jul 94* 1473-1476
- Nijmeijer, H., see Berghuis, H., *T-AC Oct 94* 2159-2162
- Nikiforuk, P.N., see Mutoh, Y., *T-AC Feb 94* 422-425
- Nikodem, M., see Olbrot, A.W., *T-AC Mar 94* 652-657
- Nikoukhah, R., see Chou, K.C., *T-AC Mar 94* 479-492
- Ocali, O., see Karan, M., *T-AC May 94* 991-995
- Olas, A. Construction of optimal Lyapunov function for systems with structured uncertainties, *T-AC Jan 94* 167-171
- Olbrot, A.W., see Sun, J., *T-AC Mar 94* 630-635
- Olbrot, A.W., and M. Nikodem Robust stabilization: some extensions of the gain margin maximization problem, *T-AC Mar 94* 652-657
- Ortega, R., see Kelly, R., *T-AC Jun 94* 1222-1224
- Ortega, R., and A. Herrera A solution to the continuous-time adaptive decoupling problem, *T-AC Aug 94* 1639-1643
- Osorio, A., see Lozano, R., *T-AC Aug 94* 1748-1751
- Ozguler, A.B., see Khargonekar, P.P., *T-AC Apr 94* 877-882
- Ozguner, I., see Schoenwald, D.A., *T-AC Aug 94* 1751-1755
- Paden, B., see Shevitz, D., *T-AC Sep 94* 1910-1914
- Pait, F.M., and A.S. Morse A cyclic switching strategy for parameter-adaptive control, *T-AC Jun 94* 1172-1183
- Pait, F.M., see Morse, A.S., *T-AC Sep 94* 1809-1818
- Pal, D., and T. Kailath Displacement structure approach to singular root distribution problems: the unit circle case, *T-AC Jan 94* 238-245
- Palaniswami, M., see Man Zhihong, *T-AC Jan 94* 154-159
- Palaniswami, M., see Gang Feng, *T-AC Nov 94* 2330-2334
- Palmer, Z.J., see Tuch, J., *T-AC Apr 94* 823-827
- Pandolfi, L., see Logemann, H., *T-AC Jan 94* 138-143
- Pankov, A.R., and A.V. Bosov Conditionally minimax algorithm for nonlinear system state estimation, *T-AC Aug 94* 1617-1620
- Pankov, A.R., see Borisov, A.V., *T-AC Dec 94* 2461-2464
- Pao, L.V., and G.F. Franklin The robustness of a proximate time-optimal controller [optional read optimal], *T-AC Sep 94* 1963-1966
- Papadakis, I.N.M., and S.C.A. Thomopoulos Binary hypothesis testing with structured adaptive networks, *T-AC Sep 94* 1967-1971
- Papavasilopoulos, G.P. Distributed algorithms with random processor failures, *T-AC May 94* 1032-1036
- Papavasilopoulos, G.P., see Li-Wen Chen, *T-AC Aug 94* 1621-1626
- Papinski, A.P., see Zhihong, M., *T-AC Dec 94* 2464-2469
- Paraskevopoulos, P.N., F.N. Koumboulis, and K.G. Tzlerakis Disturbance rejection of left-invertible generalized state space systems, *T-AC Jan 94* 185-190
- Paraskevopoulos, P.N., A.S. Tzirikos, and K.G. Arvanitis A new orthogonal series approach to state space analysis of bilinear systems, *T-AC Apr 94* 793-797
- Park, F.C. Computational aspects of the product-of-exponentials formula for robot kinematics, *T-AC Mar 94* 643-647
- Partington, J.R., and P.M. Makila Worst-case analysis of identification-BIBO robustness for closed-loop data, *T-AC Oct 94* 2171-2176
- Pasik-Duncan, B., see Duncan, T.F., *T-AC Jan 94* 118-122
- Passino, K.M., A.N. Michel, and P.J. Antsaklis Lyapunov stability of a class of discrete event systems, *T-AC Feb 94* 269-279
- Passino, K.M., A.N. Michel, and P.J. Antsaklis Correction to "Lyapunov stability of a class of discrete-event systems" (Feb 94 269-279), *T-AC Jul 94* 1531
- Patra, A., see Bose, S., *T-AC Apr 94* 803-806
- Pattipati, K.R., see Pete, A., *T-AC Aug 94* 1702-1707
- Pena, R.S.S., see Eszter, E.G., *T-AC Nov 94* 2315-2318
- Peres, P.L.D., and J.C. Geromel An alternate numerical solution to the linear quadratic problem, *T-AC Jan 94* 198-202
- Peres, P.L.D., see Geromel, J.C., *T-AC May 94* 1072-1075
- Perez, F., D. Docampo, and C. Abdallah Extreme-point robust stability results for discrete-time polynomials, *T-AC Jul 94* 1470-1472
- Perez, J.L., see Carazo, A.H., *T-AC Aug 94* 1636-1638
- Pete, A., K.R. Pattipati, and D.L. Kleinman Optimization of detection networks with multiple event structures, *T-AC Aug 94* 1702-1707
- Petersen, I.R., and D.C. McFarlane Optimal guaranteed cost control and filtering for uncertain linear systems, *T-AC Sep 94* 1971-1977
- Petr, D.W. Optimization in a class of priority-discarding policies for finite queues, *T-AC May 94* 1020-1024
- Piet-Lahanier, H., and E. Walter Bounded-error tracking of time-varying parameters, *T-AC Aug 94* 1661-1664
- Pal, K., see Poolla, K., *T-AC May 94* 951-959
- Pal, K., and D. Towsley Optimal scheduling in a machine with stochastic varying processing rate, *T-AC Sep 94* 1853-1855
- Pandya, K.S., and J. Balakrishnan Improving transient response of adaptive control systems using multiple models and switching, *T-AC Sep 94* 1861-1866
- Pandya, K.S., and J. Balakrishnan A common Lyapunov function for stable LTI systems with commuting A-matrices, *T-AC Dec 94* 2469-2471
- Pan-Touss, K., and W. Ren On the convergence of least squares estimates in white noise, *T-AC Feb 94* 364-368
- Pasian, A., see Bonivento, C., *T-AC Apr 94* 846-849
- Pann, L., see Gutman, P.-O., *T-AC Jun 94* 1268-1273

- Fimpalkhare, A.A., and B. Bandyopadhyay. Comments on "Stabilization via static output feedback"; *T-AC May 94* 1148.
- Pintelon, R., see Schoukens, J., *T-AC Aug 94* 1733-1737
- Pintelon, R., P. Guillaume, Y. Rolain, J. Schoukens, and H. Van Hamme. Parametric identification of transfer functions in the frequency domain—a survey; *T-AC Nov 94* 2245-2260
- Poh, E.-K., see Kabamba, P.T., *T-AC Jun 94* 1298-1301
- Poh, E.-K., see Kabamba, P.T., *T-AC Jul 94* 1455-1459
- Poldermann, J.W., see Weyer, E., *T-AC Aug 94* 1665-1671
- Polia, M.P., see Sun, J., *T-AC Mar 94* 630-635
- Poloujadoff, M., see Derbel, N., *T-AC Jul 94* 1429-1431
- Polyak, B.T., Ya.Z. Tsytkin, Y.Q. Shi, K.K. Yen, and C.M. Chen. Comments on "Two necessary conditions for a complex polynomial to be strictly Hurwitz and their applications in robust stability analysis" [and reply]; *T-AC May 94* 1147-1148
- Poola, K., and A. Tikku. On the time complexity of worst-case system identification; *T-AC May 94* 944-950
- Poola, K., P. Khargonekar, A. Tikku, J. Krause, and K. Nagpal. A time-domain approach to model validation; *T-AC May 94* 951-959
- Popov, V.M. Review of "Applications of Lyapunov Methods in Stability" (Halanay, A., and Rasvan, V.; 1993); *T-AC Dec 94* 2526-2527
- Pota, H.R., and T.E. Alberts. Review of "Robot Dynamics and Control" (Spong, M.W., and Vidyasagar, M.; 1989); *T-AC Aug 94* 1774-1776
- Powers, D.L., see Lovass-Nagy, V., *T-AC Jul 94* 1507-1509
- Prakash, R. Properties of a low-frequency approximation balancing method of model reduction; *T-AC May 94* 1135-1141
- Premaratne, K., and E.I. Jury. Tabular method for determining root distribution of delta-operator formulated real polynomials; *T-AC Feb 94* 352-355
- Premaratne, K., R. Salvi, N.R. Habib, and J.P. LeGall. Delta-operator formulated discrete-time approximations of continuous-time systems; *T-AC Mar 94* 581-585
- Priouret, P., see Juditsky, A., *T-AC Jun 94* 1211-1221
- Proth, J.-M., and Xiao-Lan Xie. Cycle time of stochastic event graphs: evaluation and marking optimization; *T-AC Jul 94* 1482-1486
- Pugh, A.C., N.P. Karampetakis, A.I.G. Vardoulakis, and G.E. Hayton. A fundamental notion of equivalence for linear multivariable systems; *T-AC May 94* 1141-1145

Q

- Qian-Yu Tang, see Han-Fu Chen, *T-AC Mar 94* 681-686
- Quan Wang, and Dajun Wang. A reduced-order model about structural wave control based upon the concept of degree of controllability; *T-AC Aug 94* 1711-1713
- Qui, L., and T. Chen. H_2 -optimal design of multirate sampled-data systems; *T-AC Dec 94* 2506-2511

R

- Rachid, A. Some new bounds for the spectral radius; *T-AC Jan 94* 196-198
- Radaideh, S.M., see Ahmed, N.U., *T-AC Jun 94* 1322-1326
- Radenkovic, M.S., and J.T. Bialasiewicz. Robustness of unmodified stochastic adaptive control algorithms; *T-AC Feb 94* 396-400
- Radenkovic, M.S., and D.A. Lawrence. Using burst recovery concept to obtain global stability and performance of discrete-time adaptive controller for time-varying systems; *T-AC Nov 94* 2357-2361
- Rafajlowicz, E., and W. Myszk. Efficient algorithm for a class of least squares estimation problems; *T-AC Jun 94* 1241-1243
- Ramadge, P.J., see Chong, E.K.P., *T-AC Jul 94* 1400-1410
- Rantzer, A., and A. Megretski. A convex parameterization of robustly stabilizing controllers; *T-AC Sep 94* 1802-1808
- Ravikanth, R., see Ellis, R.D., *T-AC Sep 94* 1795-1801
- Ren, W., see Nassiri-Toussi, K., *T-AC Feb 94* 364-368
- Rizzoni, G., see Jachong Park, *T-AC Jul 94* 1521-1524
- Rolain, Y., see Pintelon, R., *T-AC Nov 94* 2245-2260
- Romagnoli, J.A., see Figueroa, J.L., *T-AC Apr 94* 831-835
- Rosen, I.G., see Demetriou, M.A., *T-AC May 94* 1117-1123
- Rotstein, B. Preconditioning of transfer matrices: bounding the frequency dependent structured singular value; *T-AC Nov 94* 2287-2292
- Rotstein, H., and A. Sideris. H_∞ optimization with time-domain constraints; *T-AC Apr 94* 762-779
- Rubinstein, R.Y., see Uryas'ev, S., *T-AC Jun 94* 1263-1267
- Rugh, W.J., see Lawrence, D.A., *T-AC Nov 94* 2207-2218
- Ruiz, A.C., and H. Nijmeijer. Controllability distributions and systems approximations: a geometric approach; *T-AC Jul 94* 1473-1476
- Runolfsson, T. The equivalence between infinite-horizon optimal control of stochastic systems with exponential-of-integral performance index and stochastic differential games; *T-AC Aug 94* 1551-1563

- Rutland, N.K. The principle of matching: practical conditions for systems with inputs restricted in magnitude and rate of change; *T-AC Mar 94* 550-553
- Ryan, E.P. A nonlinear universal servomechanism; *T-AC Apr 94* 753-761

S

- Saab, S.S. On the P-type learning control; *T-AC Nov 94* 2298-2302
- Saberi, A. Review of "The H_∞ Control Problem: A State Space Approach" (Stoorvogel, A.; 1992); *T-AC Jan 94* 252-254
- Saberi, A., see Stoorvogel, A.A., *T-AC Feb 94* 355-360
- Saberi, A., see Stoorvogel, A.A., *T-AC Sep 94* 1936-1939
- Sacks, A., see Bodson, M., *T-AC Sep 94* 1939-1944
- Sain, M.K. Editorial: The 1993 George S. Axelby Outstanding Paper Award; *T-AC Feb 94* 258
- Salvi, R., see Premaratne, K., *T-AC Mar 94* 581-585
- Samad, T. Review of "Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches" (White, D.A., and Sofge, D.A., Eds.; 1992); *T-AC Jul 94* 1534-1535
- Sangchul Won, see Kwanghee Nam, *T-AC Nov 94* 2311-2314
- Sang Woo Kim, see Joon Hwa Lee, *T-AC Jan 94* 159-162
- Sastry, S., see Walsh, G., *T-AC Jan 94* 216-222
- Sayed, A.H., and T. Kailath. Extended Chandrasekhar recursions; *T-AC Mar 94* 619-623
- Sayed, A.H., T. Constantinescu, and T. Kailath. Time-variant displacement structure and interpolation problems; *T-AC May 94* 960-976
- Sayed, A.H., see Constantinescu, T., *T-AC Jul 94* 1476-1481
- Sayed, A.H., T. Kailath, and H. Lev-Ari. Generalized Chandrasekhar recursions from the generalized Schur algorithm; *T-AC Nov 94* 2265-2269
- Scattoloni, R., see Clarke, D.W., *T-AC May 94* 1052-1056
- Scherer, R., and W. Wendler. A generalization of the positive real lemma; *T-AC Apr 94* 882-886
- Schilling, R.J., see Lovass-Nagy, V., *T-AC Jul 94* 1507-1509
- Schneider, A.M., J.A. Anusiewicz, and I.S. Barghouti. Accuracy and stability of discrete-time filters generated by higher-order s-to-z mapping functions; *T-AC Feb 94* 435-441
- Schochetman, I.E., and R.L. Smith. Solution approximation in infinite horizon linear quadratic control; *T-AC Mar 94* 596-601
- Schoenwald, D.A., and I. Ozguner. Robust stabilization of nonlinear systems with parametric uncertainty; *T-AC Aug 94* 1751-1755
- Schoukens, J., and R. Pintelon. Quantifying model errors of identified transfer functions; *T-AC Aug 94* 1733-1737
- Schoukens, J., see Pintelon, R., *T-AC Nov 94* 2245-2260
- Schumacher, J.M. Review of "Controlled and Conditioned Invariants in Linear System Theory" (Basile, G., and Marro, G.; 1992); *T-AC Jan 94* 250-251
- Schwartz, H.M., see Aitken, V.C., *T-AC Sep 94* 1959-1962
- Sebek, M., see Kwakernaak, H., *T-AC Feb 94* 315-328
- Sebek, M., see Hunt, K.J., *T-AC Jul 94* 1502-1507
- Seidman, T.I. "First come, first served" can be unstable!; *T-AC Oct 94* 2166-2171
- Selbuz, H., see Eldem, V., *T-AC May 94* 1002-1006
- Seongno Lee, see Kwanghee Nam, *T-AC Nov 94* 2311-2314
- Sethi, S., and Xun Yu Zhou. Stochastic dynamic job shops and hierarchical production planning; *T-AC Oct 94* 2061-2076
- Seto, D., A.M. Annaswamy, and J. Baillieul. Adaptive control of nonlinear systems with a triangular structure; *T-AC Jul 94* 1411-1428
- Seto, D., and J. Baillieul. Control problems in super-articulated mechanical systems; *T-AC Dec 94* 2442-2453
- Sezer, M.E., and D.D. Siljak. On stability of interval matrices; *T-AC Feb 94* 368-371
- Sezer, M.E., see Karan, M., *T-AC May 94* 991-995
- Shaked, U., see Theodor, Y., *T-AC Sep 94* 1944-1948
- Shaked, U., see Theodor, Y., *T-AC Oct 94* 2130-2134
- Shamma, J.S. Robust stability with time-varying structured uncertainty; *T-AC Apr 94* 714-724
- Shapiro, A., and Y. Wardi. Nondifferentiability of the steady-state function in discrete event dynamic systems; *T-AC Aug 94* 1707-1711
- Sharifnia, A. Stability and performance of distributed production control methods based on continuous-flow models; *T-AC Apr 94* 725-737
- Shevitz, D., and B. Paden. Lyapunov stability theory of nonsmooth systems; *T-AC Sep 94* 1910-1914
- Shi, Y.Q., see Polyak, B.T., *T-AC May 94* 1147-1148
- Shieh, L.S., see Koc, C.K., *T-AC Aug 94* 1644-1647
- Shimemura, E., see Kojima, A., *T-AC Aug 94* 1694-1698
- Shimizu, K., and S. Ito. Constrained optimization in Hilbert space and a generalized dual quasi-Newton algorithm for state-constrained optimal control problems; *T-AC May 94* 982-986
- Shin-Yeu Lin. A hardware implementable receding horizon controller for constrained nonlinear systems; *T-AC Sep 94* 1893-1899

- Shir-Kuan Lin. Identification of a class of nonlinear deterministic systems with application to manipulators, *T-AC Sep 94* 1886-1893
- Shujaee, K., see Lehman, B, *T-AC Aug 94* 1673-1676
- Shuoh Rern, P T Kabamba, and D S Bernstein. Guardian map approach to robust stability of linear systems with constant real parameter uncertainty, *T-AC Jan 94* 162-164
- Siciliano, B., see Chiaverini, S, *T-AC Mar 94* 647-652
- Sideris, A., see Rotstein, H, *T-AC Apr 94* 762-779
- Sideris, A., see Sznajer, M, *T-AC Jul 94* 1497-1502
- Siljak, D.D., see Sezer, M E, *T-AC Feb 94* 368-371
- Skaar, S.B., see Baumgartner, E T, *T-AC Mar 94* 493-502
- Smith, R.L., see Schochetman, I E, *T-AC Mar 94* 596-601
- Smith, R.S., see Erickson, M A, *T-AC Sep 94* 1904-1909
- Socha, L. Some remarks on exact linearization of a class of stochastic dynamical systems, *T-AC Sep 94* 1980-1984
- Soh, Y.C., see Foo, Y K, *T-AC Jan 94* 151-154
- Song, Y.D., F L Mitchell, and H Y Lai. Control of a class of nonlinear uncertain systems via compensated inverse dynamics approach, *T-AC Sep 94* 1866-1871
- Sontag, E.D., see Sussmann, H J, *T-AC Dec 94* 2411-2425
- Soueres, P., J-Y Fourquet, and J-P Laumond. Set of reachable positions for a car, *T-AC Aug 94* 1626-1630
- Souza, S.R., see Geromel, J C, *T-AC May 94* 1072-1075
- Sparaggis, P.D., D Fowsley, and C G Cassandras. Routing with limited state information in queueing systems with blocking, *T-AC Jul 94* 1492-1497
- Speyer, J.L., see Weiss, H, *T-AC Mar 94* 540-544
- Speyer, J.L., see Banavar, R N, *T-AC Sep 94* 1914-1918
- Speyer, J.L., see Chih-Hai Fan, *T-AC Oct 94* 1986-2003
- Sreenivas, R.S. On the existence of finite state supervisors for arbitrary supervisory control problems, *T-AC Apr 94* 856-861
- Sreeram, V., P Agathoklis, and M Mansour. The generation of discrete-time q-Markov covers via inverse solution of the Lyapunov equation, *T-AC Feb 94* 381-385
- Sreeram, V. and P Agathoklis. The discrete-time q-Markov cover models with improved low-frequency approximation, *T-AC May 94* 1102-1105
- Sreeram, V. On the generalized q-Markov cover models for discrete-time systems, *T-AC Dec 94* 2502-2505
- Srikant, R. Relationship between decentralized controller design using H_∞ and stochastic risk-averse criteria, *T-AC Apr 94* 861-864
- Stankovic, S., see Musicki, D, *T-AC Jun 94* 1237-1241
- Stepanenko, Y., see Xueshan Yang, *T-AC Mar 94* 585-589
- Stepanenko, Y., see Chun-Yi Su, *T-AC Mar 94* 609-614
- Stoorvogel, A.A., A Saberi, and B M Chen. A reduced order observer based controller design for H_∞ -optimization, *T-AC Feb 94* 355-360
- Stoorvogel, A.A., and A J I M Weeren. The discrete-time Riccati equation related to the H_∞ control problem, *T-AC Mar 94* 686-691
- Stoorvogel, A.A., A Saberi, and B M Chen. The discrete-time H_∞ control problem with strictly proper measurement feedback, *T-AC Sep 94* 1936-1939
- Su, J.-H. Comments on "Stability margin evaluation for uncertain linear systems," (by C. Gong and S. Thompson, Jun 94 548-550), *T-AC Dec 94* 2523-2524
- Sugie, F., see Imura, J-I, *T-AC May 94* 1084-1089
- Sun, J., A W Olbrot, and M P Polis. Robust stabilization and robust performance using model reference control and modeling error compensation, *T-AC Mar 94* 630-635
- Sung-Joon Lee, see In-Joong Ha, *T-AC Nov 94* 2269-2274
- Sussmann, H.J., F D Sontag, and Y Yang. A general result on the stabilization of linear systems using bounded controls, *T-AC Dec 94* 2411-2425
- Syrmos, V.L., and F L Lewis. A bilinear formulation for the output feedback problem in linear systems, *T-AC Feb 94* 410-414
- Syrmos, V.L. Disturbance decoupling using constrained Sylvester equations, *T-AC Apr 94* 797-803
- Sznajer, M., see Benzaid, Z, *T-AC May 94* 1064-1066
- Sznajer, M., see Blanchini, F, *T-AC Jul 94* 1487-1492
- Sznajer, M., and A Sideris. Feedback control of quantized constrained systems with applications to neuromorphic controllers design, *T-AC Jul 94* 1497-1502
- Sznajer, M. An exact solution to general SISO mixed H_2/H_∞ problems via convex optimization, *T-AC Dec 94* 2511-2517
- Taylor, D.G., see Khayatian, A, *T-AC Jun 94* 1292-1297
- Taylor, D.G., see Khayatian, A, *T-AC Sep 94* 1848-1852
- Te-Jen Su, and Wen-Jye Shyr. Robust D-stability for linear uncertain discrete time-delay systems, *T-AC Feb 94* 425-428
- Tempo, R., see Holot, C V, *T-AC Feb 94* 391-396
- Te-Ping Tsai, see Yen-Ting Hsu, *T-AC Aug 94* 1722-1726
- Tesi, A., A Vicino, and G Zappa. Convexity properties of polynomials with assigned root location, *T-AC Mar 94* 668-672
- Theodor, Y., and U Shaked. Game theory approach to H_∞ -optimal discrete-time fixed-point and fixed-lag smoothing, *T-AC Sep 94* 1944-1948
- Theodor, Y., and U Shaked. H_∞ multiple objective robust controllers for infinite-horizon single measurement single control input problems, *T-AC Oct 94* 2130-2134
- Thomopoulos, S.C.A., see Papadakis, I N M, *T-AC Sep 94* 1967-1971
- Thompson, S., see Gong, C, *T-AC Mar 94* 548-550
- Tian-Shen Tang. On the textured iterative algorithms for a class of tridiagonal linear equations, *T-AC Mar 94* 592-596
- Tian Yu-Ping, Feng Chun-Bo, and Xin Xin. Robust stability of polynomials with multilinearly dependent coefficient perturbations, *T-AC Mar 94* 554-558
- Tikku, A., see Poola, K, *T-AC May 94* 944-950
- Tikku, A., see Poola, K, *T-AC May 94* 951-959
- Tilbury, D., see Walsh, G, *T-AC Jan 94* 216-222
- Tomei, P. Tracking control of flexible joint robots with uncertain parameters and disturbances, *T-AC May 94* 1067-1072
- Tomei, P., see Marino, R, *T-AC Jun 94* 1314-1319
- Tongwen Chen. On stability robustness of a dual-rate control system, *T-AC Jan 94* 164-167
- Tonielli, A., see Bonivento, C, *T-AC Apr 94* 846-849
- Torres, J., see Lozano, R, *T-AC Aug 94* 1748-1751
- Towsley, D., see Sparaggis, P D, *T-AC Jul 94* 1492-1497
- Towsley, D., see Nain, P, *T-AC Sep 94* 1853-1855
- Treil, S. A counterexample on continuous coprime factors, *T-AC Jun 94* 1262-1263
- Trinh, H., and M Aldeen. On the stability of linear systems with delayed perturbations, *T-AC Sep 94* 1948-1951
- Tropearsky, M.T., see Gonzalez, G A, *T-AC Oct 94* 2145-2148
- Tsai, W.K., see Antonio, J K, *T-AC Feb 94* 385-391
- Tsai, W.K., Gang Huang, and Wei Lu. Fast parallel recursive aggregation methods for simulation of dynamical systems, *T-AC Mar 94* 534-540
- Tsai, W.K., see Antonio, J K, *T-AC Sep 94* 1839-1844
- Tseng, P., see Zhi-Quan Luo, *T-AC May 94* 1123-1129
- Tsinias, J., and N Kalouptsidis. A correction note on "Output feedback stabilization", *T-AC Apr 94* 806
- Tsinias, J., see Kazakos, D, *T-AC Oct 94* 2111-2113
- Tsirikos, A.S., see Paraskevopoulos, P N, *T-AC Apr 94* 793-797
- Tsitsiklis, J.N., see Irving, W W, *T-AC Apr 94* 835-838
- Tsyypkin, Ya.Z., see Polyak, B T, *T-AC May 94* 1147-1148
- Tuch, J., A Feuer, and Z J Palmor. Time delay estimation in continuous linear time-invariant systems, *T-AC Apr 94* 823-827
- Turan, D., L L Mingori, and G C Goodwin. Loop transfer recovery design using biased and unbiased controllers, *T-AC Mar 94* 601-605
- Tzicerakis, K.G., see Paraskevopoulos, P N, *T-AC Jan 94* 185-190

U

- Uchida, K., see Kojima, A, *T-AC Aug 94* 1694-1698
- Uetake, Y. Adaptive observer for continuous descriptor systems, *T-AC Oct 94* 2095-2100
- Uryas'ev, S., and R Y Rubinstein. On relaxation algorithms in computation of noncooperative equilibria, *T-AC Jun 94* 1263-1267

V

- Vadari, S.V., see Aggoune, M E, *T-AC Oct 94* 2106-2110
- Vakil, P., see Jian-Qiang Hu, *T-AC Sep 94* 1875-1880
- Van Hamme, H., see Pintelon, R, *T-AC Nov 94* 2245-2260
- van Schuppen, J.H. Tuning of Gaussian stochastic control systems, *T-AC Nov 94* 2178-2190
- Vardoulakis, A.I.G., see Pugh, A C, *T-AC May 94* 1141-1145
- Verghese, G.C., see Karl, W C, *T-AC Jan 94* 222-226
- Verma, M.S., see Hunt, L R, *T-AC Oct 94* 2113-2118
- Verriest, E.L., see Lehman, B, *T-AC May 94* 898-912
- Viaro, U., see Krajewski, W, *T-AC Oct 94* 2126-2129
- Viaro, U., see Beghi, A, *T-AC Dec 94* 2494-2496
- Vicino, A., see Tesi, A, *T-AC Mar 94* 668-672
- Villani, L., see Chiaverini, S, *T-AC Mar 94* 647-652
- Voulgaris, P. Control of asynchronous sampled data systems, *T-AC Jul 94* 1451-1455

T

- Tijne, M., M M'Saad, and L Dugard. Discrete-time compensators with loop transfer recovery, *T-AC Jun 94* 1259-1262
- Tsam Kang, see Dohyoung Chung, *T-AC Aug 94* 1698-1702
- Tsouris, S., and C Burgat. Positively invariant sets for constrained continuous-time systems with cone properties, *T-AC Feb 94* 401-405
- Tsouris, S. Corrections on "Positively invariant sets for constrained continuous-time systems with cone properties" (Feb 94 401-405), *T-AC Aug 94* 1771

W

- Wahlberg, B. System identification using Kautz models; *T-AC Jun 94* 1276-1282
- Walsh, G., D. Tilbury, S. Sastry, R. Murray, and J.P. Laumond. Stabilization of trajectories for systems with nonholonomic constraints; *T-AC Jan 94* 216-222
- Walter, E., and L. Jaulin. Guaranteed characterization of stability domains via set inversion; *T-AC Apr 94* 886-889
- Walter, E., see Piet-Lahanier, H., *T-AC Aug 94* 1661-1664
- Wang, L., and W.R. Cluett. Optimal choice of time-scaling factor for linear system approximations using Laguerre models; *T-AC Jul 94* 1463-1467
- Wang, L.-S., see Lian, K.-Y., *T-AC Dec 94* 2426-2441
- Wang, M.-Z., see Duan, G.-R., *T-AC Dec 94* 2490-2494
- Wang, Q., see Weiss, H., *T-AC Mar 94* 540-544
- Wang, Q.G., see Lee, T.H., *T-AC Mar 94* 565-568
- Wang, Y.W., see Haddad, W.M., *T-AC Apr 94* 827-831
- Wang, Y.W., W.M. Haddad, and D.S. Bernstein. Robust strong stabilization via modified Popov controller synthesis; *T-AC Nov 94* 2284-2287
- Ward, Y., see Shapiro, A., *T-AC Aug 94* 1707-1711
- Watson, L.T., see Ge, Y., *T-AC Jun 94* 1302-1305
- Weeren, A.J.T.M., see Stoorvogel, A.A., *T-AC Mar 94* 686-691
- Wei-Bing Gao, see Feng Zheng, *T-AC Aug 94* 1714-1718
- Weijian Zhang. The spectral density of a nonlinear damping model: multi-DOF case; *T-AC Feb 94* 406-410
- Wei Lin, see Byrnes, C.I., *T-AC Jan 94* 83-98
- Wei Lin, and C.I. Byrnes. Design of discrete-time nonlinear control systems via smooth feedback; *T-AC Nov 94* 2340-2346
- Wei Lu, see Tsai, W.K., *T-AC Mar 94* 534-540
- Weinert, H.L., see Jie Chen, *T-AC May 94* 1114-1116
- Wei-qian Sun, P.P. Khargonekar, and Duksun Shim. Solution to the positive real control problem for linear time-invariant systems; *T-AC Oct 94* 2034-2046
- Wei Ren, and P.R. Kumar. Stochastic adaptive prediction and model reference control; *T-AC Oct 94* 2047-2060
- Weiss, H., Q. Wang, and J.L. Speyer. System characterization of positive real conditions; *T-AC Mar 94* 540-544
- Weiss, M. Spectral and inner-outer factorizations through the constrained Riccati equation; *T-AC Mar 94* 677-681
- Wei-Yong Yan, B.D.O. Anderson, and R.R. Bitmead. On the gain margin improvement using dynamic compensation based on generalized sampled-data hold functions; *T-AC Nov 94* 2347-2354
- Weller, S.R., and G.C. Goodwin. Hysteresis switching adaptive control of linear multivariable systems; *T-AC Jul 94* 1360-1375
- Wendler, W., see Scherer, R., *T-AC Apr 94* 882-886
- Wen-June Wang, Gin-Hol Wu, and Dah-Ching Yang. Variable structure control design for uncertain discrete-time systems; *T-AC Jan 94* 99-102
- Wen-Jye Shyr, see Te-Jen Su, *T-AC Feb 94* 425-428
- Weyer, E., I.M.Y. Marcelis, and J.W. Poldermann. Limitations of robust adaptive pole placement control; *T-AC Aug 94* 1665-1671
- Wiberg, D.M., see Kanellakopoulos, I., *T-AC Aug 94* 1776
- Wigren, T. Convergence analysis of recursive identification algorithms based on the nonlinear Wiener model; *T-AC Nov 94* 2191-2206
- Willsky, A.S., see Chou, K.C., *T-AC Mar 94* 464-478
- Willsky, A.S., see Chou, K.C., *T-AC Mar 94* 479-492
- Wimmer, H.K. Consistency of a pair of generalized Sylvester equations; *T-AC May 94* 1014-1016
- Wong, E.W.M., and T.-S.P. Yum. The optimal multicopy Aloha; *T-AC Jun 94* 1233-1236
- Wonham, W.M., see Brandin, B.A., *T-AC Feb 94* 329-342
- Wonham, W.M., see Yong Li, *T-AC Mar 94* 512-531
- Wonham, W.M., see Li, Y., *T-AC Aug 94* 1771
- Wook Hyun Kwon, see Joon Hwa Lee, *T-AC Jan 94* 159-162
- Wu, H.R., see Zhihong, M., *T-AC Dec 94* 2464-2469
- Wu, J.W., and Keum-Shik Hong. Delay-independent exponential stability criteria for time-varying discrete delay systems; *T-AC Apr 94* 811-814

X

- Xianggang Yu, see Hsu, C.S., *T-AC Aug 94* 1679-1681
- Xiao-Hui Zhao, see Lozano, R., *T-AC Jan 94* 47-58
- Xiao-Lan Xie. Superposition properties and performance bounds of stochastic timed-event graphs; *T-AC Jul 94* 1376-1386
- Xiao-Lan Xie, see Proth, J.-M., *T-AC Jul 94* 1482-1486
- Xiaoping Yun, see Yamamoto, Y., *T-AC Jun 94* 1326-1332
- Xiangzhi Ji, see Lin-Zhang Lu, *T-AC Aug 94* 1682-1685
- Xin Xin, see Tian Yu-Ping, *T-AC Mar 94* 554-558
- Xi-Ren Ao, see Dye-Jyun Ma, *T-AC Jul 94* 1460-1463
- Xi Sun, and Ji Feng Zhang. Adaptive stabilization of bilinear systems; *T-AC Jan 94* 207-211
- Xu, B. Correction to "An improved Razumikhin-type theorem and its application" (Apr 94 839-841); *T-AC Nov 94* 2368

- Xu, S.J., see Darouach, M., *T-AC Mar 94* 606-609
- Xueshan Yang, and Y. Stepanenko. A stability criterion for discrete nonlinear systems with time delayed feedback; *T-AC Mar 94* 585-589
- Xun Yu Zhou, see Sethi, S., *T-AC Oct 94* 2061-2076

Y

- Yamada, T., and S. Kataoka. On some LP problems for performance evaluation of timed marked graphs; *T-AC Mar 94* 696-698
- Yamamoto, Y. A function space approach to sampled data control systems and tracking problems; *T-AC Apr 94* 703-713
- Yamamoto, Y., and Xiaoping Yun. Coordinating locomotion and manipulation of a mobile manipulator; *T-AC Jun 94* 1326-1332
- Yamamoto, Y., see Hayakawa, Y., *T-AC Nov 94* 2278-2284
- Yang, Y., see Sussmann, H.J., *T-AC Dec 94* 2411-2425
- Yang Chengwu, see Zou Yun, *T-AC Jul 94* 1436-1439
- Yangchun Zhong, see Campbell, S.L., *T-AC May 94* 1047-1052
- Yangsheng Xu, see Ju-Jang Lee, *T-AC Feb 94* 414-419
- Yau-Tarng Juang, see Yen-Ting Hsu, *T-AC Aug 94* 1722-1726
- Yen, K.K., see Polyak, B.T., *T-AC May 94* 1147-1148
- Yeng Chai Soh, see Lihua Xie, *T-AC Jun 94* 1310-1314
- Yen-Ting Hsu, Te-Ping Tsai, and Yau-Tarng Juang. The equivalence of model-following systems designed in the time domain and frequency domain; *T-AC Aug 94* 1722-1726
- Yong Li, and W.M. Wonham. Control of vector discrete-event systems II. Controller synthesis; *T-AC Mar 94* 512-531
- Yongqing Liu, see Bugong Xu, *T-AC Apr 94* 839-841
- Yoshikawa, T., see Imura, J.-I., *T-AC May 94* 1084-1089
- Young, P.M., see Braatz, R.P., *T-AC May 94* 1000-1002
- Young Man Cho, Guanghan Xu, and T. Kailath. Fast identification of state-space models via exploitation of displacement structure; *T-AC Oct 94* 2004-2017
- Yu-Chi Ho. Heuristics, rules of thumb, and the 80/20 proposition; *T-AC May 94* 1025-1027
- Yuksel, O., see Afacan, T., *T-AC Nov 94* 2261-2262
- Yum, T.-S.P., see Wong, E.W.M., *T-AC Jun 94* 1233-1236
- Yung-Chun Wu, and Nie-Zen Yen. An optimal algorithm for sampled-data robust servomechanism controller using exponential hold; *T-AC Jan 94* 112-117
- Yung-Chun Wu, and Nie-Zen Yen. A ripple free sampled-data robust servomechanism controller using exponential hold; *T-AC Jun 94* 1287-1291
- Yu Tang, and M.A. Arteaga. Adaptive control of robot manipulators based on passivity; *T-AC Sep 94* 1871-1875

Z

- Zames, G., L. Lin, and Le Yi Wang. Fast identification n-widths and uncertainty principles for LTI and slowly varying systems; *T-AC Sep 94* 1827-1838
- Zanasi, R., see Bonivento, C., *T-AC Apr 94* 846-849
- Zappa, G., see Tesi, A., *T-AC Mar 94* 668-672
- Zasadzinski, M., see Darouach, M., *T-AC Mar 94* 606-609
- Zasadzinski, M., see Darouach, M., *T-AC Aug 94* 1755-1758
- Zaslavsky, B. Stabilization of oscillations by a nonnegative feedback control; *T-AC Jun 94* 1273-1276
- Zeitouni, O., see Dembo, A., *T-AC Feb 94* 360-364
- Zelentsovsky, A.L. Nonquadratic Lyapunov functions for robust stability analysis of linear uncertain systems; *T-AC Jan 94* 135-138
- Zeungnam Bien, see Jin-Hoon Kim, *T-AC Jan 94* 202-207
- Zexiang Li, see Fernandes, C., *T-AC Mar 94* 450-463
- Zhabko, A.P., see Kharitonov, V.L., *T-AC Dec 94* 2388-2397
- Zhihong, M., A.P. Paplinski, and H.R. Wu. A robust MIMO terminal sliding mode control scheme for rigid robotic manipulators; *T-AC Dec 94* 2464-2469
- Zhibin Qu, J.F. Dorsey, and D.M. Dawson. Model reference robust control of a class of SISO systems; *T-AC Nov 94* 2219-2234
- Zhi-Quan Luo, and P. Tseng. On the rate of convergence of a distributed asynchronous routing algorithm; *T-AC May 94* 1123-1129
- Zhi Yu Zhang, and F. Kozin. On almost sure sample stability of nonlinear stochastic dynamic systems; *T-AC Mar 94* 560-565
- Zhongzhi Hu. Decentralized stabilization of large scale interconnected systems with delays; *T-AC Jan 94* 180-182
- Zhou, K., see Doyle, J., *T-AC Aug 94* 1575-1587
- Zhuquan Zang, and R.R. Bitmead. Transient bounds for adaptive control systems; *T-AC Jan 94* 171-175
- Zigang Pan, and T. Basar. H^∞ -optimal control for singularly perturbed systems. II. Imperfect state measurements; *T-AC Feb 94* 280-299
- Zou Yun, and Yang Chengwu. An algorithm for the computation of all eigenvalues; *T-AC Jul 94* 1436-1439

SUBJECT INDEX

A

absolute stability

convex approach, absol. stabil. problem. *Gapski, P.B.*, +, *T-AC Sep 94* 1929-1932

dyn. lin. time-delay systs., robust controller design. *Mahmoud, M.S.*, +, *T-AC May 94* 995-999

nonlin. differential-delay systs., absol. stabil. *Gil, M.I.*, *T-AC Dec 94* 2481-2484

parametrically uncertain nonlin. feedback systs., absol. stabil. *Marquez, H.J.*, +, *T-AC Mar 94* 664-668

C generator excitation; cf. Power generation control, excitation ctuators

uncertain SISO min. phase lin. syst., nonlin. universal servomechanism. *Ryan, E.P.*, *T-AC Apr 94* 753-761

adaptive control

bilinear systs., adaptive stabilization, least sq. *Xi Sun*, +, *T-AC Jan 94* 207-211

book review; Adaptive Control Systems (Isermann, R., et al.; 1992). *Kanellakopoulos, I.*, +, *T-AC Aug 94* 1776

book review; Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches (White, D.A., and Sofge, D.A., Eds.; 1992). *Samad, T.*, *T-AC Jul 94* 1534-1535

car reachable posns. computation. *Soueres, P.*, +, *T-AC Aug 94* 1626-1630

chaotic behavior in adaptive control systs. *Gonzalez, G.A.*, +, *T-AC Oct 94* 2145-2148

complex discrete-time control systs., simul. automation. *Ellis, R.D.*, +, *T-AC Sep 94* 1795-1801

continuous descriptor systs., adaptive observer. *Uetake, Y.*, *T-AC Oct 94* 2095-2100

continuous direct adaptive control, saturation input constraint. *Cishen Zhang*, +, *T-AC Aug 94* 1718-1722

continuous-time adaptive decoupling control design. *Ortega, R.*, +, *T-AC Aug 94* 1639-1643

continuous-time nonlin. systs. adaptive control by neural nets. *Fu-Chuang Chen*, +, *T-AC Jun 94* 1306-1310

corrections to "On the stability proof of adaptive schemes with static normalizing signal and parameter projection" (Jan 93 170-173). *Ioannou, P.*, *T-AC Apr 94* 896

cyclicly switched param.-adaptive control systs., MIMO design models/internal regulators. *Morse, A.S.*, +, *T-AC Sep 94* 1809-1818

discrete-time adaptive nonlin. syst., least sq. estimator. *Kanellakopoulos, I.*, *T-AC Nov 94* 2362-2365

discrete time-varying systs., robust adaptive controller. *Changyun Wen*, *T-AC May 94* 987-991

dyn. systs., robust adaptive control design. *Khorasani, K.*, *T-AC Aug 94* 1726-1732

failure detect./isolation/accommodation syst. *Chia-Chi Tsui*, *T-AC Nov 94* 2318-2321

feedback lin. systs. adaptive tracking. *Marino, R.*, +, *T-AC Jun 94* 1314-1319

first-order nonlin. syst. adaptive control. *Brogliato, B.*, +, *T-AC Aug 94* 1764-1768

Gaussian stochastic control syst. tuning. *van Schuppen, J.H.*, *T-AC Nov 94* 2178-2190

harmonic generation in adaptive feedforward cancellation schemes. *Bodson, M.*, +, *T-AC Sep 94* 1939-1944

interconnected systs., decentralized adaptive regulation. *Changyun Wen*, *T-AC Oct 94* 2163-2166

jump-Markov systs., stabilizing control law. *Dufour, F.*, +, *T-AC Nov 94* 2354-2357

lin. systs., nonlin. design of adaptive controllers. *Krstic, M.*, +, *T-AC Apr 94* 738-752

multiple models/switching approach. *Narendra, K.S.*, +, *T-AC Sep 94* 1861-1866

nonlin. design of adaptive controllers. *Krstic, M.*, +, *T-AC Apr 94* 738-752

nonlin. systs., triangular struct., adaptive control. *Seto, D.*, +, *T-AC Jul 94* 1411-1428

nonlin. uncertain systs. tracking control. *Song, Y.D.*, +, *T-AC Sep 94* 1866-1871

nonminimum phase first-order continuous-time systs., adaptive stabilization. *Lozano, R.*, +, *T-AC Aug 94* 1748-1751

parab. systs. direct adaptive control. *Keum Shik Hong*, +, *T-AC Oct 94* 2018-2033

param.-adaptive control, cyclic switching strategy. *Pail, F.M.*, +, *T-AC Jun 94* 1172-1183

pole placement without excitation probing sigs. *Lozano, R.*, +, *T-AC Jan 94* 47-58

predictive controller, robustness props. *Clarke, D.W.*, +, *T-AC May 94* 1052-1056

rigid spacecraft adaptive attitude control. *Egeland, O.*, +, *T-AC Apr 94* 842-846

robot adaptive control based on passivity. *Yu Tang*, +, *T-AC Sep 94* 1871-1875

robot controller, adaptive, Lyapunov stabil. *Egeland, O.*, +, *T-AC Aug 94* 1671-1673

robot manipulators, globally convergent adaptive controller. *Mahmoud, M.S.*, +, *T-AC Jan 94* 148-151

robust adaptive pole placement control. *Weyer, E.*, +, *T-AC Aug 94* 1665-1671

robust adaptive regulation with min. prior knowledge, comment. *Ji Feng Zhang*, *T-AC Mar 94* 605

robust approach. *Gang Feng*, *T-AC Aug 94* 1738-1742

robust direct adaptive controllers, normalization tech. *Gang Feng*, +, *T-AC Nov 94* 2330-2334

robustness and multitone instabilities. *Bodson, M.*, *T-AC Apr 94* 864-870

sing.-free multivariable MRAC. *Moctezuma, R.G.*, +, *T-AC Sep 94* 1856-1860

single-arm dyn., robust variable struct./and switching- Σ adaptive control. *Li-Wen Chen*, +, *T-AC Aug 94* 1621-1626

SISO LTI discrete-time systs., param.-adaptive controller. *Kreisselmeier, G.*, *T-AC Sep 94* 1819-1826

slowly varying and LTI systs., ident. and uncertainty principles. *Zames, G.*, +, *T-AC Sep 94* 1827-1838

stabilization, adaptive, nonlin. time-varying controller. *Miller, D.E.*, *T-AC Jul 94* 1347-1359

stochastic adaptive control algms., unmodified, robustness. *Radenkovic, M.S.*, +, *T-AC Feb 94* 396-400

stochastic adaptive prediction/MRAC. *Wei Ren*, +, *T-AC Oct 94* 2047-2060

time-invariant lin. syst. adaptive control. *Karason, S.P.*, +, *T-AC Nov 94* 2325-2330

time varying systs., discrete time adaptive controller, global stabil. *Radenkovic, M.S.*, +, *T-AC Nov 94* 2357-2361

transient bounds. *Zhuquan Zang*, +, *T-AC Jan 94* 171-175

uncertain SISO min. phase lin. syst., nonlin. universal servomechanism. *Ryan, E.P.*, *T-AC Apr 94* 753-761

Adaptive control; cf. Model reference adaptive control

Adaptive estimation

distributed param. systs., adaptive estim., persistence of excitation. *Demetriou, M.A.*, +, *T-AC May 94* 1117-1123

generalized Chandrasekhar recursions from generalized Schur algm. *Sayed, A.H.*, +, *T-AC Nov 94* 2265-2269

ident., signed output error adaptive, convergence. *Garnett, J.*, +, *T-AC Jul 94* 1387-1399

Adaptive estimation; cf. Adaptive Kalman filtering; Adaptive observers

Adaptive filters

generalized Chandrasekhar recursions from generalized Schur algm. *Sayed, A.H.*, +, *T-AC Nov 94* 2265-2269

time-variant displacement struct. and interpolation problems. *Sayed, A.H.*, +, *T-AC May 94* 960-976

Adaptive Kalman filtering

extended Chandrasekhar recursions. *Sayed, A.H.*, +, *T-AC Mar 94* 619-623

Adaptive observers

continuous descriptor systs., adaptive observer. *Uetake, Y.*, *T-AC Oct 94* 2095-2100

Adaptive signal processing; cf. Adaptive filters

Adaptive systems

global optim., adaptive partitioned random search. *Bo Tang, Z.*, *T-AC Nov 94* 2235-2244

least sq. estim. in white noise, convergence. *Nassiri-Toussi, K.*, +, *T-AC Feb 94* 364-368

Aerospace control

attitude control problem, comment. *Fjellstad, O.-E.*, +, *T-AC Mar 94* 699-700

Algebra

control algms., automatic differentiation appl. *Campbell, S.L.*, +, *T-AC May 94* 1047-1052

jump-Markov systs., stabilizing control law. *Dufour, F.*, +, *T-AC Nov 94* 2354-2357

output feedback problem in lin. systs., bilinear formulation. *Syrmos, V.L.*, +, *T-AC Feb 94* 410-414

sing. perturbed control systs. and nonlin. differential- $\text{alg. eqns. Krishnan, H.}, +, *T-AC May 94* 1079-1084$

three-block generalized/std. Riccati eqns. comparison. *Darouach, M.*, +, *T-AC Aug 94* 1755-1758

Algebra; cf. Polynomials; Set theory; Vectors

AM

intersample props., robustness, and sensitivity, quantitat./qualitat. anal. *Feuer, A.*, +, *T-AC May 94* 1042-1047

Approximation methods

- delay systs approx, Laguerre formula *Lam, J* T-AC Jul 94 1517-1521
 delta-operator formulated discrete-time approx *Premaratne K.* + T-AC Mar 94 581-585
 discrete-time systs, mixed H_2/H_∞ control, exact convex optim based soln *Senaier, M* T-AC Dec 94 2511-2517
 H_∞ model reduction computational scheme *Kavranoglu D* T-AC Jul 94 1447-1451
 H_∞ -norm approx of systs by const matrices *Kavranoglu D* T-AC May 94 1006-1009
 infinite horizon LQ control, soln approx *Schochetman I E* + T-AC Mar 94 596-601
 * interval systs order reduction, Routh-Pade approx *Bandyopadhyay B* + T-AC Dec 94 2454-2456
 model reduction, LF approx balancing, props *Prakash R* T-AC May 94 1135-1141
 nonlin syst controllability distribts / systs approx *Ruiz A C* + T-AC Jul 94 1473-1476
 rational transfer fn, optimal L_∞ approx *Kavranoglu D* + T-AC Sep 94 1899-1904
 textured iter algms for tridiagonal lin eqns *Tian-Shen Tang* T-AC Mar 94 592-596
Approximation methods; cf. Interpolation, Linear approximation, Minimax methods, Polynomial approximation, Stochastic approximation
ARMA processes; cf. Autoregressive moving-average processes
Array processing
 state-space models fast ident *Young Man Cho* + T-AC Oct 94 2004-2017
Artificial intelligence, cf. Knowledge-based systems
Asymptotic control
 observers for nonlin systs in steady state *Hunt L R* + T-AC Oct 94 2113-2118
Asymptotic stability
 approx feedback lin, stabil *Kwanghee Nam* + T-AC Nov 94 2311-2314
 asymptotic stabil criteria for lin systs time delay *Juing-Huei Su* + T-AC Jun 94 1341-1344
 common Lyapunov fn, stable LTI systs, commuting A-matrices *Narendra K S* + T-AC Dec 94 2469-2471
 corrections to "An improved Razumikhin-type theorem and its application" (Apr 94 839-841) *Xu B* T-AC Nov 94 2368
 delay-independent exponential stabil criteria for time-varying discrete delay systs *Wu J W* + T-AC Apr 94 811-814
 discrete-time nonlin control systs design via smooth feedback *Wei Lin* + T-AC Nov 94 2340-2346
 discrete-time syst exponential stabil, observer design *Aitken V C* + T-AC Sep 94 1959-1962
 dyn systs, robust adaptive control design *Khorasani K* T-AC Aug 94 1726-1732
 extended horizon predictive control *Fanyin Kong* + T-AC Jul 94 1467-1470
 flexible joint robots global regulation, approx diff *Kelly R* + T-AC Jun 94 1222-1224
 interval LTI systs stabilization *Kehui Wei* T-AC Jan 94 22-32
 logical discrete event systs, Lyapunov stabil *Passino K M* + T-AC Feb 94 269-279
 LQG controllers, stable, weighting and covariance matrices *Halevi Y* T-AC Oct 94 2104-2106
 multivariable systs, quantitat robustness measures *Hong-Giou Chen* + T-AC Apr 94 807-810
 nonlin regulator, num design *Kreisselmeier G* + T-AC Jan 94 33-46
 nonlin systs, asymptotic model matching *Di Benedetto M D* + T-AC Aug 94 1539-1550
 nonminimum phase first-order continuous-time systs, adaptive stabilization *Lozano R* + T-AC Aug 94 1748-1751
 observers for nonlin systs in steady state *Hunt L R* + T-AC Oct 94 2113-2118
 pos real systs nonlin controllers *Bernstein D S* + T-AC Jul 94 1513-1517
 Razumikhin-type theorem, improved, appls *Bugong Xu* + T-AC Apr 94 839-841
 robot manipulators, compliant, force/posn regulation *Chiaverini S* + T-AC Mar 94 647-652
 robust strong stabilization via modified Popov controller synthesis *Wang Y W*, + T-AC Nov 94 2284-2287
 stabil of lin systs, delayed perturb *Trinh H* + T-AC Sep 94 1948-1951
 stick-slip avoidance, PD control *Dupont P E* T-AC May 94 1094-1097
 uncertain systs, asymptotic stabil region estim *Han Ho Choi* + T-AC Nov 94 2275-2278
Attitude control
 comment and error correction *Fjellstad O-E* + T-AC Mar 94 699-700
 rigid spacecraft adaptive attitude control *Egeland, O* + T-AC Apr 94 842-846

Automata

- logical discrete event systs, Lyapunov stabil *Passino K M* + T-AC Feb 94 269-279
 timed discrete-event systs, supervisory control *Brandin B A* + T-AC Feb 94 329-342
Automata; cf. Finite automata
Autoregressive moving-average processes
 CARMA plants, stabilizing I-O receding horizon control *Chisci L* + T-AC Mar 94 614-618
 stochastic adaptive prediction/MRAC *Wei Ren* + T-AC Oct 94 2047-2060
Awards
 1993 George S Axelby Outstanding Paper Award given to I Kannellakopoulos, P V Kokotovic, and A S Morse *Sain M K* T-AC Feb 94 258

B**Bayes procedures**

- sens fusion, decentralized detect optimal thresholds *Irving W W* + T-AC Apr 94 835-838

Bilinear systems

- bilinear systs, adaptive stabilization, least sq *Xi Sun* + T-AC Jan 94 207-211
 bilinear systs, state space anal orthogonal series approach *Paraskevopoulos P N* + T-AC Apr 94 793-797
 finite-time optimal control of bilinear systs successive approx procedure *Aganovic Z* + T-AC Sep 94 1932-1935
 output feedback problem in lin systs, bilinear formulation *Syrmos V I* + T-AC Feb 94 410-414

Biological control systems

- cat, falling near-optimal nonholonomic motion planning *Fernandes C* + T-AC Mar 94 450-463

Biomechanics

- cat, falling near-optimal nonholonomic motion planning *Fernandes C* + T-AC Mar 94 450-463

Book reviews

- Adaptive Control Systems (Isermann, R, et al 1992) *Kanellakopoulos I* + T-AC Aug 94 1776
 Applications of Lyapunov Methods in Stability (Halanay A and Rasvan V, 1993) *Popov I M* T-AC Dec 94 2526-2527
 Applied Optimal Control and Estimation (Lewis I I 1992) *Kamen E W* T-AC Aug 94 1773-1774
 Controlled and Conditioned Invariants in Linear System Theory (Basile G, and Marro, G, 1992) *Schumacher J M* T-AC Jan 94 250-251
 Feedback Control Theory (Doyle, J et al 1992) *Moore K L* T-AC Jul 94 1532-1534
 Handbook of Intelligent Control Neural Fuzzy and Adaptive Approaches (White D A, and Sofge, D A Eds 1992) *Samad T* T-AC Jul 94 1534-1535
 H_∞ Control Problem A State Space Approach (Stoorvogel A, 1992) *Saberi A* T-AC Jan 94 252-254
 Linear Multivariable Control Algebraic Analysis and Synthesis Methods (Vardulakis, A I G, 1991) *Lewis F L* T-AC Jul 94 1536
 Linear System Theory (Rugh W J 1993) *Khalil H K* T-AC Dec 94 2528-2529
 New Tools for Robustness of Linear Systems (Barmish B R 1994) T-AC Dec 94 2525-2526
 Nonlinear System Analysis, 2nd edn (Vidyasagar M 1993) *Abed E H* T-AC Jul 94 1535-1536
 Nonlinear Systems (Khalil H 1992) *Grizzle J W* T-AC Jan 94 251-252
 Robot Dynamics and Control (Spong, M W and Vidyasagar, M, 1989) *Pota H R* + T-AC Aug 94 1774-1776

Boundary-value problems

- dyn syst simul, fast parallel recursive aggregation *Fu W K* + T-AC Mar 94 534-540
 multipoint boundary value stochastic systs stationarity/reciprocity *Ji Chen* + T-AC May 94 1114-1116

Brushless rotating machines

- brushless DC motors, feedback-lin control charactn *In-Joong Ha* + T-AC Mar 94 673-677

C**Cascade systems**

- cascaded nonlin systs, global robust stabilization *Imura J-I* + T-AC May 94 1084-1089
 lin continuous-time systs, global stabilization, bounded control *Sussmann H J* + T-AC Dec 94 2411-2425
 robust control design, cascade struct approach *Bonivento C* + T-AC Apr 94 846-849

- stabilization of discrete-time nonlin sys^ts, global *Byrnes C J* + *T-AC Jan 94* 83-98
- Chaos**
adaptive control sys^ts, chaotic behavior *Gonzalez G A* + *T-AC Oct 94* 2145-2148
- Circuit noise; cf. Filter noise**
- Circuits; cf. Sample-and-hold circuits**
- Communication channels**
optimal multicopy Aloha *Wong E W M* + *T-AC Jun 94* 1233-1236
- Communication network routing**
data network path formulated optimal routing time complexity *Antonio J K* + *T-AC Feb 94* 385-391
deterministic/stochastic queuing networks anal *Cheng-Shang Chang T-AC May 94* 913-931
distributed asynchronous routing, convergence rate *Zhi-Quan Luo* + *T-AC May 94* 1123-1129
queuing networks anal, projection techs *Mooi Choo Chuah T-AC Aug 94* 1588-1599
- Communication switching, cf. Packet switching**
- Communication systems; cf. Data communication**
- Communication system software**
distributed asynchronous routing convergence rate *Zhi-Quan Luo* + *T-AC May 94* 1123-1129
- Communication traffic**
deterministic/stochastic queuing networks anal *Cheng-Shang Chang T-AC May 94* 913-931
optimal multicopy Aloha *Wong E W M* + *T-AC Jun 94* 1233-1236
queuing networks anal, projection techs *Mooi Choo Chuah T-AC Aug 94* 1588-1599
- Compensation**
CARMA plants, stabilizing I-O receding horizon control *Chisci L* + *I-AC Mar 94* 614-618
continuous control sys^t, digital equiv design methods comparison *Hall S R* *T-AC Feb 94* 420-421
continuous time nonminimum phase lin sys^ts gain margin improvement *Wei Yong Yan* + *I-AC Nov 94* 2347-2354
diagonal decoupling dyn output feedback/const precompensator *Eldem V* *T-AC Mar 94* 503-511
discrete-time compensators, IIR *Tadjine M* + *I-AC Jun 94* 1259-1262
discrete-time uncertain sys^ts ultimate boundedness control Lyapunov *Blanchini F* *T-AC Feb 94* 428-433
H_∞ compensator design min order observers *Hsu C S* + *T-AC Aug 94* 1679-1681
H_∞-optim reduced order observer based controller *Stoorvogel A A* + *T-AC Feb 94* 355-360
input-output decoupling nonlin interactor appl *Di Benedetto M D* + *T-AC Jun 94* 1246-1250
nonlin feedback parametrically uncertain sys^ts absol stabil *Marquez H J* + *T-AC Mar 94* 664-668
nonlin sys^ts asymptotic model matching *Di Benedetto M D* + *T-AC Aug 94* 1539-1550
nonlin sys^ts asymptotic tracking necessary conditions *Grizzle J W* + *T-AC Sep 94* 1782-1794
nonlin uncertain sys^ts tracking control *Song Y D* + *T-AC Sep 94* 1866-1871
rational l¹ suboptimal compensators for continuous-time sys^ts *Blanchini F* + *I-AC Jul 94* 1487-1492
rigid robotic manipulators, robust tracking control *Man Zhihong* + *T-AC Jan 94* 154-159
robust stabilization/perform, MRAC modeling error compensation *Sun J* + *T-AC Mar 94* 630-635
robust strong stabilization via modified Popov controller synthesis *Wang Y W* + *T-AC Nov 94* 2284-2287
ISO discrete-time sys^ts preview tracking *Halpern M E* *T-AC Mar 94* 589-592
Smith predictor for controlling proc, integrator and long dead-time *Astrom K J* + *T-AC Feb 94* 343-345
state observer/feedback, compensators, comment *Bender D J* *T-AC Feb 94* 447-448
uncertain nonlin sys^ts, struct invariance *Castro-Linares R* + *T-AC Oct 94* 2154-2158
ubr sys^ts control *Karl W C* + *T-AC Jan 94* 222-226
- Complexity theory**
discrete-time filters from high-order s-to-z mappings *Schneider A M*, + *T-AC Feb 94* 435-441
dyn sys^t simul, fast parallel recursive aggregation *Tsai W K* + *T-AC Mar 94* 534-540
heuristics, rules of thumb, and the 80/20 proposition *Yu-Chi Ho* *T-AC May 94* 1025-1027
integrated probabilistic data assoc *Musicki D* + *T-AC Jun 94* 1237-1241
l¹ sq estim, efficient algm *Rafajlowicz E* + *T-AC Jun 94* 1241-1243
- robust anal, uncertainty value sets *Eszter E G*, + *T-AC Nov 94* 2315-2318
single-machine scheduling, resource optimal control *Cheng T C E*, + *T-AC Jun 94* 1243-1246
worst case ident, H[∞] model validation *Guoxiang Gu*, *T-AC Aug 94* 1657-1661
worst-case sys^t ident, time complexity *Poolla K*, + *T-AC May 94* 944-950
μ calc, computational complexity *Braatz R P* + *T-AC May 94* 1000-1002
- Computational geometry**
nonlin sys^t controllability distrib^s/sys^ts approx *Ruiz A C* + *T-AC Jul 94* 1473-1476
- Computation time**
data network path formulated optimal routing, time complexity *Antonio J K* + *T-AC Feb 94* 385-391
path formulated optimal routing algm, time complexity *Antonio J K* + *T-AC Sep 94* 1839-1844
- Computer communication; cf. Data communication**
- Computer instructions**
quantized constrained sys^ts, feedback control, appls to neuromorphic controllers design *Sznajder M* + *T-AC Jul 94* 1497-1502
- Computers; cf. Distributed computing, Parallel processing**
- Continuous-time systems**
constrained continuous-time sys^ts with cone props, pos invariant sets *Tarbouriech S* + *T-AC Feb 94* 401-405
continuous direct adaptive control, saturation input constraint *Cishen Zhang* + *T-AC Aug 94* 1718-1722
continuous-time adaptive decoupling control design *Ortega R* + *T-AC Aug 94* 1639-1643
continuous time nonminimum phase lin sys^ts, gain margin improvement *Wei-Yong Yan* + *T-AC Nov 94* 2347-2354
continuous time sys^ts, transfer fns, model errors *Schoukens J* + *T-AC Aug 94* 1733-1737
corrections to "On undershoot in SISO systems" (Mar 94 578-581) *de la Barra S B A I* *T-AC Aug 94* 1771
corrections to "Positively invariant sets for constrained continuous-time systems with cone properties" (Feb 94 401-405) *Tarbouriech S* *T-AC Aug 94* 1771
deadbeat ripple-free tracking *Jetto L* *T-AC Aug 94* 1759-1764
delta-operator formulated discrete-time approx *Premaratne K* + *T-AC Mar 94* 581-585
descriptor sys^ts, lin feedback closed-loop Stackelberg strategy *Hua Xu* + *T-AC May 94* 1097-1102
lin continuous-time sys^ts, global stabilization, bounded controls *Sussmann H J* + *T-AC Dec 94* 2411-2425
nonminimum phase first-order continuous-time sys^ts, adaptive stabilization *Lozano R* + *T-AC Aug 94* 1748-1751
perturbed lin sys^ts upper covariance bounds *Bolzern P* + *T-AC Mar 94* 623-626
quadratic stabilization of continuous time sys^ts *Mahmoud M S* + *T-AC Oct 94* 2135-2139
rational l¹ suboptimal compensators for continuous-time sys^ts *Blanchini F* + *T-AC Jul 94* 1487-1492
SISO sys^ts, step response, undershoot *Leon de la Barra S B A* *T-AC Mar 94* 578-581
time delay estim in continuous LTI sys^ts *Tuch J* + *T-AC Apr 94* 823-827
uncertain continuous-time implicit sys^ts, regional pole placement, robustness *Chun-Hsiung Fang* + *T-AC Nov 94* 2303-2307
- Controllability**
adaptive control sys^ts, transient bounds *Zhuquan Zang* + *T-AC Jan 94* 171-175
arbitrary supervisory control, finite state supervisors *Sreenivas R S* *T-AC Apr 94* 856-861
cat falling, near-optimal nonholonomic motion planning, appl *Fernandes C*, + *T-AC Mar 94* 450-463
discrete-time q-Markov cover generation, inverse Lyapunov method *Sreeram V* + *T-AC Feb 94* 381-385
feedback control by multirate PWM *Khayatian A* + *T-AC Jun 94* 1292-1297
impulse differential lin sys^ts, constrained controllability *Benzaid Z* + *T-AC May 94* 1064-1066
infinite horizon LQ control, soln approx *Schochetman I E* + *T-AC Mar 94* 596-601
multirate sampled-data sys^ts struct props *Longhi S* *T-AC Mar 94* 692-696
multiscale sys^ts, Kalman filters, Riccati eqns *Chou K C*, + *T-AC Mar 94* 479-492
multivariable nonlin controller vibr damping *Kanestrom R K*, + *T-AC Sep 94* 1925-1928

- nonlin. syst. controllability distrib./systs. approx. Ruiz, A.C., +, T-AC Jul 94 1473-1476
- pos. realness conditions, charactn. assuming controllability. Weiss, H., +, T-AC Mar 94 540-544
- quantized constrained systs., feedback control, appls. to neuromorphic controllers design Senaler, M., +, T-AC Jul 94 1497-1502
- rigid spacecraft systs. controllability, central gravitational field. Lian, K.-Y., +, T-AC Dec 94 2426-2441
- SISO LTI discrete-time systs., param.-adaptive controller. Kreisselmeier, G., T-AC Sep 94 1819-1826
- state deadbeat control problem, general soln. Eldem, V., +, T-AC May 94 1002-1006
- structural wave control, reduced-order model. Quan Wang, +, T-AC Aug 94 1711-1713
- supervisory control, blocking and controllability of Petri nets. Giua, A., +, T-AC Apr 94 818-823
- switched-mode power converters, multirate modeling and control design Khayatian, A., +, T-AC Sep 94 1848-1852
- timed discrete-event systs., supervisory control. Brandin, B.A., +, T-AC Feb 94 329-342
- time-invariant lin. systs. controllability/observability. Kaming Wang, +, T-AC Jul 94 1443-1447
- Control systems**
- matching principle for systs., restricted inputs Rutland, N.K., T-AC Mar 94 550-553
- Control systems; cf. Specific topic**
- Convergence of numerical methods**
- 2D stochastic approx., convergence/diff. eqn limit Dye-Jyun Ma, +, T-AC Jul 94 1439-1442
- data network path formulated optimal routing, time complexity. Antonio, J.K., +, T-AC Feb 94 385-391
- H ∞ control syst ident., least sq methods, comment Livstone, M.M., +, T-AC Jul 94 1531
- noncooperative equilibria computation, relax. algms Uryas'ev, S., +, T-AC Jun 94 1263-1267
- parab. systs. direct adaptive control. Keum Shik Hong, +, T-AC Oct 94 2018-2033
- path formulated optimal routing algm., time complexity Antonio, J.K., +, T-AC Sep 94 1839-1844
- recursive ident., nonlin. Wiener model. Wigren, T., T-AC Nov 94 2191-2206
- sing.-free multivariable MRAC. Moclezuma, R.G., +, T-AC Sep 94 1856-1860
- Cost-optimal control**
- failure prone mfg. syst. hedging point policies. Jian-Qiang Hu, +, T-AC Sep 94 1875-1880
- filtering/optimal cost control, uncertain lin. systs Petersen, I.R., +, T-AC Sep 94 1971-1977
- Covariance analysis**
- perturbed lin. systs., upper covariance bounds Bolzern, P., +, T-AC Mar 94 623-626
- Covariance matrices**
- lin. discrete-time systs., optimum lin. recursive estim Carazo, A.H., +, T-AC Aug 94 1636-1638
- LQG controllers, stable, weighting and covariance matrices. Halevi, Y., T-AC Oct 94 2104-2106
- nonminimum phase first-order continuous-time systs., adaptive stabilization. Lozano, R., +, T-AC Aug 94 1748-1751
- time-varying params., bounded error ident. Bittanti, S., +, T-AC May 94 1106-1110
- D**
- Data communication**
- distributed asynchronous routing, convergence rate. Zhi-Quan Luo, +, T-AC May 94 1123-1129
- network path formulated optimal routing, time complexity. Antonio, J.K., +, T-AC Feb 94 385-391
- Data transmission; cf. Data communication**
- DC motors**
- brushless DC motors, feedback-lin. control charactn In-Joong Ha, +, T-AC Mar 94 673-677
- LQ optimal regulators, stabil. robustness Dohyoung Chung, +, T-AC Aug 94 1698-1702
- ripple free sampled-data robust servomechanism controller, exponential hold. Yung-Chun Wu, +, T-AC Jun 94 1287-1291
- Decision-making**
- 2D stochastic approx., convergence/diff. eqn. limit. Dye-Jyun Ma, +, T-AC Jul 94 1439-1442
- heuristics, rules of thumb, and the 80/20 proposition Yu-Chi Ho, T-AC May 94 1025-1027
- risk-averse decentralized discrete-time LEQG games. Srikant, R., T-AC Apr 94 861-864
- Decision-making; cf. Bayes procedures; Signal detection**
- Decoupling of systems**
- continuous-time adaptive decoupling control design. Ortega, R., +, T-AC Aug 94 1639-1643
- diagonal decoupling, dyn. output feedback/const. precompensator. Eldem, V., T-AC Mar 94 503-511
- disturbance decoupling, constrained Sylvester eqns. Syrmos, V.L., T-AC Apr 94 797-803
- input-output decoupling, nonlin. interactor appl. Di Benedetto, M.D., +, T-AC Jun 94 1246-1250
- input-output lin., state equivalence and decoupling. In-Joong Ha, +, T-AC Nov 94 2269-2274
- row-by-row stable decoupling, static state feedback, struct. soln. Martinez Garcia, J.C., +, T-AC Dec 94 2457-2460
- Delay estimation**
- continuous LTI systs. Tuch, J., +, T-AC Apr 94 823-827
- Delay systems**
- contact stabil. of simple posn. controllers, effect of time delay and discrete control. Fiala, J., +, T-AC Apr 94 870-873
- corrections to "An improved Razumikhin-type theorem and its application" (Apr 94 839-841). Xu, B., T-AC Nov 94 2368
- DEDS, observability, delay Bose, S., +, T-AC Apr 94 803-806
- delay-independent exponential stabil. criteria for time-varying discrete delay systs. Wu, J.W., +, T-AC Apr 94 811-814
- delay systs. approx., Laguerre formula Lam, J., T-AC Jul 94 1517-1521
- distributed delay lin. systs., feedback stabilisation Feng Zheng, +, T-AC Aug 94 1714-1718
- D-stabil., robust, for lin. uncertain discrete delay systs. Te-Jen Su, +, T-AC Feb 94 425-428
- dyn. lin. time-delay systs., robust controller design Mahmoud, M.S., +, T-AC May 94 995-999
- H ∞ -optimal discrete-time fixed-point and fixed-lag smoothing, game theory approach Theodor, Y., +, T-AC Sep 94 1944-1948
- lin. syst. approx., time-scaling factor, Laguerre models Wang, L., +, T-AC Jul 94 1463-1467
- nonlin. differential-delay systs., absol. stabil. Gil, M.I., T-AC Dec 94 2481-2484
- quadratic stabilization of continuous time systs. Mahmoud, M.S., +, T-AC Oct 94 2135-2139
- Razumikhin-type theorem, improved, appls. Bugong Xu, +, T-AC Apr 94 839-841
- robust stabilization of syst., control delays Kojima, A., +, T-AC Aug 94 1694-1698
- robust stabil., quasipolynomial convex directions/testing sets Kharitonov V.L., +, T-AC Dec 94 2388-2397
- stabil. of lin. systs., delayed perturb. Trinh, H., +, T-AC Sep 94 1948-1951
- state delayed systs., memoryless H ∞ controllers Joon Hwa Lee, +, T-AC Jan 94 159-162
- time delay estim. in continuous LTI systs. Tuch, J., +, T-AC Apr 94 823-827
- time lag nonlin. systs., vibr. control. Lehman, B., +, T-AC May 94 898-912
- uncertain lin. systs., delay depend., robust stabil. Bugong Xu, T-AC Nov 94 2365
- uncertain nonlin. interconnected systs., time-varying state delay stabilizing control Mahmoud, M.S., T-AC Dec 94 2484-2488
- Descriptor systems; cf. Singular systems**
- Detection; cf. Fault diagnosis; Signal detection**
- Diagnosis; cf. Fault diagnosis**
- Difference equations**
- three-block generalized/std. Riccati eqns. comparison Darouach, M., +, T-AC Aug 94 1755-1758
- Differentiability**
- control algms., automatic differentiation appl. Campbell, S.L., +, T-AC May 94 1047-1052
- flexible joint robots, global regulation, approx. diff. Kelly, R., +, T-AC Jun 94 1222-1224
- recursive prediction error algm. log likelihood fn. derivatives. Hool, M.A., T-AC Mar 94 662-664
- Differential equations**
- 2D stochastic approx., convergence/diff. eqn. limit. Dye-Jyun Ma, T-AC Jul 94 1439-1442
- control algms., automatic differentiation appl. Campbell, S.L., +, T-AC May 94 1047-1052
- finite-time optimal control of bilinear systs., successive approx. proced. Aganovic, Z., +, T-AC Sep 94 1932-1935
- impulse differential lin. systs., constrained controllability. Benzaid, Z., T-AC May 94 1064-1066
- lin. multivariable systs., fund. notion of equivalence. Pugh, A.C., +, T-AC May 94 1141-1145

- Lyapunov stabil. theory of nonsmooth systs *Shevitz, D.*, +, *T-AC Sep 94* 1910-1914
- polynomial differential eqns., domain of attraction estim. *Levin, A.*, *T-AC Dec 94* 2471-2475
- time lag nonlin. systs., vibr control *Lehman, B.*, +, *T-AC May 94* 898-912
- time-varying functional diff. eqns., delay independent stabil conditions/decay estim *Lehman, B.*, +, *T-AC Aug 94* 1673-1676
- Differential equations; cf. Nonlinear differential equations; Partial differential equations; Riccati equations, differential; Stochastic differential equations
- Differential geometry
freq. response arcs, stable polynomial, convexity *Keqin Gu*, *T-AC Nov 94* 2262-2265
- Diffusion processes
almost sure sample stabil of nonlin stochastic dyn systs *Zhi Yu Zhang*, +, *T-AC Mar 94* 560-565
- Digital control
continuous control syst., digital equiv design methods comparison *Hall, S R.*, *T-AC Feb 94* 420-421
- Digital filters; cf. Recursive digital filters
- Digital image processing; cf. Image processing
- Discrete-event systems
arbitrary supervisory control, finite state supervisors *Sreenivas, R S.*, *T-AC Apr 94* 856-861
- corrections to "Control of vector discrete-event systems I - The base model" (Aug 93 1214-1227) *Li, Y.*, +, *T-AC Aug 94* 1771
- corrections to "Lyapunov stability of a class of discrete-event systems" (Feb 94 269-279) *Passino, K M.*, +, *T-AC Jul 94* 1531
- DEDS, observability, delay *Bose, S.*, +, *T-AC Apr 94* 803-806
- DEDS, steady state fn, nondifferentiability *Shapiro, A.*, +, *T-AC Aug 94* 1707-1711
- discrete event systs, decentralized control, normality theorem *Lin, F.*, +, *T-AC May 94* 1089-1093
- logical discrete event systs, Lyapunov stabil *Passino, K M.*, +, *T-AC Feb 94* 269-279
- supervisory control, blocking and controllability of Petri nets *Giua, A.*, +, *T-AC Apr 94* 818-823
- timed discrete-event systs, supervisory control *Brandin, B A.*, +, *T-AC Feb 94* 329-342
- variable lookahead supervisory control, state inform *Hadj-Alouane, N B.*, +, *T-AC Dec 94* 2398-2410
- vector discrete-event syst controller synthesis *Yong Li*, +, *T-AC Mar 94* 512-531
- Discrete-time filters
discrete-time filters from high-order s-to-z mappings *Schneider, A M.*, +, *T-AC Feb 94* 435-441
- Discrete-time systems
adaptive control systs, chaotic behavior *Gonzalez, G A.*, +, *T-AC Oct 94* 2145-2148
- adaptive nonlin syst, least-sq estimator *Kanellakopoulos, I.*, *T-AC Nov 94* 2362-2365
- adaptive predictive controller, robustness anal *Clarke, D W.*, +, *T-AC May 94* 1052-1056
- asynchronous sampled data syst control, design *Voulgaris, P.*, *T-AC Jul 94* 1451-1455
- block multirate input-output model for sampled-data control systs *Jakubowski, A M.*, +, *T-AC May 94* 1145-1147
- certainty equivalence control, dyn games *James, M R.*, *T-AC Nov 94* 2321-2324
- compensators, LTR *Tadjine, M.*, +, *T-AC Jun 94* 1259-1262
- complex discrete-time control systs., simul automation *Ellis, R D.*, +, *T-AC Sep 94* 1795-1801
- contact stabil of simple posn controllers, effect of time delay and discrete control *Fiala, J.*, +, *T-AC Apr 94* 870-873
- continuous control syst, digital equiv design methods comparison *Hall, S R.*, *T-AC Feb 94* 420-421
- continuous time nonminimum phase lin systs, gain margin improvement *Wei-Yong Yan*, +, *T-AC Nov 94* 2347-2354
- corrections to "Robust stability and performance via fixed-order dynamic compensation: The discrete-time case" (May 93 776-782) *Haddad, W M.*, +, *T-AC Aug 94* 1772
- deadbeat ripple-free tracking *Jetto, L.*, *T-AC Aug 94* 1759-1764
- DEDS, steady state fn, nondifferentiability *Shapiro, A.*, +, *T-AC Aug 94* 1707-1711
- delay-independent exponential stabil criteria for time-varying discrete delay systs *Wu, J W.*, +, *T-AC Apr 94* 811-814
- delta-operator formulated discrete-time approx. *Premaratne, K.*, +, *T-AC Mar 94* 581-585
- delta-operator formulated real polynomials, tabular method for determining root distrib *Premaratne, K.*, +, *T-AC Feb 94* 352-355
- δ-stabil., robust, for lin. uncertain discrete delay systs. *Te-Jen Su*, +, *T-AC Feb 94* 425-428
- dual-rate control syst., sampled-data, stabil. robustness. *Tongwen Chen*, *T-AC Jan 94* 164-167
- dyn. syst. simul., fast parallel recursive aggregation. *Tsal, W.K.*, +, *T-AC Mar 94* 534-540
- exponential stabil., observer design. *Aitken, V.C.*, +, *T-AC Sep 94* 1959-1962
- extreme-point robust stabil., discrete-time polynomials *Perez, F.*, +, *T-AC Jul 94* 1470-1472
- feedback control by multirate PWM *Khayatian, A.*, +, *T-AC Jun 94* 1292-1297
- fn space approach. *Yamamoto, Y.*, *T-AC Apr 94* 703-713
- generalized q-Markov cover models *Sreeram, V.*, *T-AC Dec 94* 2502-2505
- Hammerstein syst., polynomial ident *Lang Zi-Qiang*, *T-AC Mar 94* 569-573
- H_∞ control of discrete-time uncertain systs *Geromel, J C.*, +, *T-AC May 94* 1072-1075
- H_∞ control problem, strictly proper meas feedback *Stoorvogel, A A.*, +, *T-AC Sep 94* 1936-1939
- H_∞ -optimal discrete-time fixed-point and fixed-lag smoothing, game theory approach *Theodor, Y.*, +, *T-AC Sep 94* 1944-1948
- H_∞ type problem *Hayakawa, Y.*, +, *T-AC Nov 94* 2278-2284
- intersample props, robustness, and sensitivity, quantitat/qualitat anal. *Feuer, A.*, +, *T-AC May 94* 1042-1047
- Kalman filtering for uncertain discrete-time systs *Lihua Xie*, +, *T-AC Jun 94* 1310-1314
- lin discrete syst pole assignment. *Benzaouia, A.*, *T-AC Oct 94* 2091-2095
- lin discrete time syst, perturb., comment *Eslami, M.*, *T-AC Aug 94* 1768-1769
- lin discrete-time systs, optimum lin recursive estim. *Carazo, A H.*, +, *T-AC Aug 94* 1636-1638
- lin systs., pos invariant sets *De Santis, E.*, *T-AC Jan 94* 245-249
- LTI discrete-time systs, robust stabil anal *Karan, M.*, +, *T-AC May 94* 991-995
- Markovian jump lin systs, lin. min MSE estim *Costa, O L V.*, *T-AC Aug 94* 1685-1689
- MIMO stochastic discrete syst state estim *Liu Danyang*, +, *T-AC Oct 94* 2087-2091
- mixed H_2/H_∞ control, exact convex optim based soln *Sznaier, M.*, *T-AC Dec 94* 2511-2517
- model matching, suboptimal perfect, with noise, MRACS *Muloh, Y.*, +, *T-AC Feb 94* 422-425
- multichannel gain margin improvement, sampled-data hold fns *Chang Yang*, +, *T-AC Mar 94* 657-661
- multirate sampled-data systs, H_2 -optimal design *Qui, L.*, +, *T-AC Dec 94* 2506-2511
- multirate sampled-data systs struct props *Longhi, S.*, *T-AC Mar 94* 692-696
- multivariable nonminimum phase discrete-time systs LTR procedure. *Leon de la Barra S, B A.*, *T-AC Mar 94* 574-577
- nonlin control systs, design via smooth state feedback *Wei Lin*, +, *T-AC Nov 94* 2340-2346
- nonlin discrete-time systs, input-to-state stabil condition and global stabilization *Kazakos, D.*, +, *T-AC Oct 94* 2111-2113
- nonlin syst control, stabil property *Jie Huang*, +, *T-AC Nov 94* 2307-2311
- nonlin systs., stabil, time delayed feedback *Xueshan Yang*, +, *T-AC Mar 94* 585-589
- nonlin syst state estim, conditionally min algm *Pankov, A R.*, +, *T-AC Aug 94* 1617-1620
- optimal dyn output feedback for nonzero set point regulation, discrete-time case *Haddad, W M.*, +, *T-AC Sep 94* 1921-1925
- optimal filtering, stochastic discrete-time systs, unknown inputs *Borisov, A V.*, +, *T-AC Dec 94* 2461-2464
- partially obs discrete-time nonlin systs, risk-sensitive control and dyn. games *James, M R.*, +, *T-AC Apr 94* 780-792
- periodically time-varying lin discrete-time plants, decentralized control *Khargonekar, P P.*, +, *T-AC Apr 94* 877-882
- periodic discrete-time Riccati eqn, num soln *Hench, J J.*, +, *T-AC Jun 94* 1197-1210
- pos of dyn systs, nonpositive coeff matrices *d'Alessandro, P.*, +, *T-AC Jan 94* 131-134
- q-Markov cover generation, inverse Lyapunov method *Sreeram, V.*, +, *T-AC Feb 94* 381-385
- q-Markov cover models *Sreeram, V.*, +, *T-AC May 94* 1102-1105
- range tracking loops, large deviation anal. *Dembo, A.*, +, *T-AC Feb 94* 360-364
- rational L^1 suboptimal compensators for continuous-time systs *Blanchini, F.*, +, *T-AC Jul 94* 1487-1492
- ripple free sampled-data robust servomechanism controller, exponential hold. *Yung-Chun Wu*, +, *T-AC Jun 94* 1287-1291
- risk-averse decentralized discrete-time LEQG games *Srikant, R.*, *T-AC Apr 94* 861-864

† Check author entry for coauthors

† Check author entry for subsequent corrections/comments

- robust adaptive pole placement control. *Weyer, E.*, +, *T-AC Aug 94* 1665-1671
- robust control design, cascade struct. approach. *Bonivento, C.*, +, *T-AC Apr 94* 846-849
- robust servomech. control, exponential hold, optimal algm. *Yung-Chun Wu*, +, *T-AC Jan 94* 112-117
- sampled data control sys., H_∞ type problem. *Hayakawa, Y.*, +, *T-AC Nov 94* 2278-2284
- sampled-data sys., wordlength constraint, stabil./perform. *Fialho, I.J.*, +, *T-AC Dec 94* 2476-2481
- SISO discrete-time sys. preview tracking. *Halpern, M.E.*, *T-AC Mar 94* 589-592
- SISO LTI discrete-time sys., param.-adaptive controller. *Kreisselmeier, G.*, *T-AC Sep 94* 1819-1826
- stabil. domains, assigned root location polynomials, convexity props. *Tesi, A.*, +, *T-AC Mar 94* 668-672
- stabilization of discrete-time nonlin sys., global. *Byrnes, C.I.*, +, *T-AC Jan 94* 83-98
- stochastic control, lin.-exponential-Gaussian. *Chih-Hai Fan*, +, *T-AC Oct 94* 1986-2003
- stochastic lin. discrete sys., state estim. algm. *Ahmed, M.S.*, *T-AC Aug 94* 1652-1656
- time varying sys., discrete time adaptive controller, global stabil. *Radenkovic, M.S.*, +, *T-AC Nov 94* 2357-2361
- time-varying sys., robust adaptive controller. *Changyun Wen*, *T-AC May 94* 987-991
- uncertain discrete-time sys., variable struct. control design. *Myszkowski, P.*, +, *T-AC Nov 94* 2366-2367
- uncertain sys., sampled-data controller design. *Dolphus, R.M.*, *T-AC May 94* 1036-1042
- uncertain sys., ultimate boundedness control, Lyapunov. *Blanchini, F.*, *T-AC Feb 94* 428-433
- unstable plants, sampled-data observers, generalized holds. *Haddad, W.M.*, +, *T-AC Jan 94* 229-234
- VSS control design for uncertain discrete-time sys. *Wen-June Wang*, +, *T-AC Jan 94* 99-102
- Distributed computing**
- distributed asynchronous routing, convergence rate. *Zhi-Quan Luo*, +, *T-AC May 94* 1123-1129
- random proc. failures and distributed algm. *Papavassilopoulos, G.P.*, *T-AC May 94* 1032-1036
- Distributed control**
- decentralized stable factors and parameterization of decentralized controllers. *Date, R.*, +, *T-AC Feb 94* 347-351
- eigenstructure assignment by decentralized output feedback. *Guang-Ren Duan*, *T-AC May 94* 1009-1014
- large scale interconnected sys. with delays, decentralized stabilization. *Zhongzhi Hu*, *T-AC Jan 94* 180-182
- multivariable syst. decentralized control, closed-loop props. *Campo, P.J.*, +, *T-AC May 94* 932-943
- periodically time-varying lin. discrete-time plants, decentralized control. *Khargonekar, P.P.*, +, *T-AC Apr 94* 877-882
- prod. control methods, distributed, stabil. and perform. *Sharifnia, A.*, *T-AC Apr 94* 725-737
- risk-averse decentralized discrete-time LEQG games. *Srikant, R.*, *T-AC Apr 94* 861-864
- Distributed decision-making**
- detect. networks with multiple event struct., optim. *Pete, A.*, +, *T-AC Aug 94* 1702-1707
- Distributed estimation**
- multiresolutional distributed filtering. *Lang Hong*, *T-AC Apr 94* 853-856
- Distributed-parameter systems**
- adaptive estim., persistence of excitation. *Demetriou, M.A.*, +, *T-AC May 94* 1117-1123
- evol. eqns., parab. partial DE, exponential stabil. *Keum-Shik Hong*, +, *T-AC Jul 94* 1432-1436
- multibody flexible sys., elastic mode estim. *Karray, F.*, +, *T-AC May 94* 1016-1020
- multi-DOF nonlin. damping model, spectral dens. *Weijian Zhang*, *T-AC Feb 94* 406-410
- parab. sys. direct adaptive control. *Keum Shik Hong*, +, *T-AC Oct 94* 2018-2033
- SISO-distributed plants, optimal mixed sensitivity. *Flamm, D.S.*, +, *T-AC Jun 94* 1150-1165
- Duality**
- guaranteed param. estim. problem with uncertain stats., Kalman-Bucy filter accuracy. *Matasov, A.I.*, *T-AC Mar 94* 635-639
- lin. min.-phase plants, stabilizing controllers parameterization. *Glaria, J.J.*, +, *T-AC Feb 94* 433-434
- LP, timed marked graphs eval. *Yamada, T.*, +, *T-AC Mar 94* 696-698
- multirate sampled-data sys. struct. props. *Longhi, S.*, *T-AC Mar 94* 692-696
- observers, lin., parameterization and design. *Ding, X.*, +, *T-AC Aug 94* 1648-1652
- uncertain sys., stabilizing control design, quasiconvex optim. *Keqin Gu*, *T-AC Jan 94* 127-131
- Dynamic programming**
- finite queues, priority-discarding policies. *Petr, D.W.*, *T-AC May 94* 1020-1024
- partially obs. discrete-time nonlin. sys., risk-sensitive control and dyn. games. *James, M.R.*, +, *T-AC Apr 94* 780-792
- prod. control, multiple time scales approach. *Jiang, J.*, +, *T-AC Nov 94* 2292-2297
- risk-averse decentralized discrete-time LEQG games. *Srikant, R.*, *T-AC Apr 94* 861-864
- E**
- Eigenstructure assignment**
- eigenstructure assignment by decentralized output feedback. *Guang-Ren Duan*, *T-AC May 94* 1009-1014
- fault detect. filters, eigenstructure assignment. *Jaehong Park*, +, *T-AC Jul 94* 1521-1524
- observer design, eigenvalue assignment, residual generation. *Magni, J.-F.*, +, *T-AC Feb 94* 441-447
- output feedback problem in lin. sys., bilinear formulation. *Syrmos, V.L.*, +, *T-AC Feb 94* 410-414
- Eigenstructure assignment; cf. Pole assignment; Zero assignment**
- Eigenvalues/eigenfunctions**
- 2D general discrete state-space models, eigenvalues calc. *Zou Yun*, +, *T-AC Jul 94* 1436-1439
- continuous alg. Riccati eqn. eigenvalues. *Komaroff, N.*, *T-AC Mar 94* 532-534
- distributed delay lin. sys., feedback stabilisation. *Feng Zheng*, +, *T-AC Aug 94* 1714-1718
- eigenstructure assignment by decentralized output feedback. *Guang-Ren Duan*, *T-AC May 94* 1009-1014
- fault detect. filters, eigenstructure assignment. *Jaehong Park*, +, *T-AC Jul 94* 1521-1524
- interval LTI sys., stabilization. *Kehui Wei*, *T-AC Jan 94* 22-32
- large sparse Lyapunov eqns., approx. soln. *Gudmundsson, T.*, +, *T-AC May 94* 1110-1114
- nonnegative feedback control, oscills. stabilisation. *Zaslavsky, B.*, *T-AC Jun 94* 1273-1276
- observer design, eigenvalue assignment, residual generation. *Magni, J.-F.*, +, *T-AC Feb 94* 441-447
- output feedback problem in lin. sys., bilinear formulation. *Syrmos, V.L.*, +, *T-AC Feb 94* 410-414
- prod. control methods, distributed, stabil. and perform. *Sharifnia, A.*, *T-AC Apr 94* 725-737
- Riccati eqn., discrete-time alg., closed loop eigenvalues. *Lin-Zhang Lu*, +, *T-AC Aug 94* 1682-1685
- spectral syst. finite-dimens. approx. *Erickson, M.A.*, +, *T-AC Sep 94* 1904-1909
- structurally uncertain sys., optimal Lyapunov fns. *Olas, A.*, *T-AC Jan 94* 167-171
- tridiagonal symmetric interval matrices, eigenvalues. *Commercon, J.C.*, *T-AC Feb 94* 377-379
- uncertain continuous-time implicit sys., regional pole placement, robustness. *Chun-Hsiung Fang*, +, *T-AC Nov 94* 2303-2307
- unknown input observers design. *Darouach, M.*, *T-AC Mar 94* 698-699
- Error analysis**
- bounded-error tracking of time-varying params. *Piet-Lahanier, H.*, +, *T-AC Aug 94* 1661-1664
- MRAC perform. anal./improvement, new tracking error criteria. *Dattani, A.*, +, *T-AC Dec 94* 2370-2387
- random param. tracking, robust algm. *Juditsky, A.*, +, *T-AC Jun 94* 1211-1221
- recursive prediction error algm. log likelihood fn. derivatives. *Hook, M.A.*, *T-AC Mar 94* 662-664
- time-varying params., bounded error ident. *Bittanti, S.*, +, *T-AC May 94* 1106-1110
- Estimation**
- book review; Applied Optimal Control and Estimation" (Lewis, F., 1992). *Kamen, E.W.*, *T-AC Aug 94* 1773-1774
- model struct. selection test, instrumental variable, statist. props. *Nguyen, Nghia Duong*, +, *T-AC Jan 94* 211-215
- nonlin. syst. H_∞ control, output-feedback based. *Lu, W.-M.*, +, *T-AC Jan 94* 2517-2524
- polynomial differential eqns., domain of attraction estim. *Levin, A.*, *T-AC Dec 94* 2471-2475

Estimation; cf. Delay estimation, Filtering, Innovations methods (stochastic processes), Maximum-likelihood estimation, Mean-square-error methods, Nonparametric estimation, Parameter estimation, Recursive estimation, Smoothing methods

F

Failure analysis

mfg syst, failure prone hedging point *Jian-Qiang Hu* + *T-AC Sep 94* 1875-1880

Failure analysis; cf. Fault diagnosis, Reliability**Fault diagnosis**

failure detect /isolation/accommodation syst *Chia-Chi Tsui* *T-AC Nov 94* 2318-2321

filters fault detect type eigenstructure assignment *Jaehong Park* + *T-AC Jul 94* 1521-1524

Fault tolerance

failure detect /isolation/accommodation syst *Chia-Chi Tsui* *T-AC Nov 94* 2318-2321

Feedback systems

2D syst model of Roesser, decomp *Afacan T* + *T-AC Nov 94* 2261-2262

approx feedback lin, stabil *Kwanghee Nam* + *T-AC Nov 94* 2311-2314

attitude control problem, comment *Fjellstad O E* + *T-AC Mar 94* 699-700

book review, Feedback Control Theory (Doyle J et al, 1992) *Moore A L* *T-AC Jul 94* 1532-1534

brushless DC motors, feedback-lin control charactn *In Joong Ha* + *T-AC Mar 94* 673-677

deterministic nonlin syst ident *Shir Kuan Lin* *T-AC Sep 94* 1886-1893

floating platform modeling and control *Damen A A H* + *T-AC May 94* 1075-1078

Gaussian stochastic control syst tuning *van Schuppen J H* *T-AC Nov 94* 2178-2190

H₂ multiojective robust control infinite-horizon *Iheodor Y* + *T-AC Oct 94* 2130-2134

jump-Markov systs stabilizing control law *Dufour F* + *T-AC Nov 94* 2354-2357

MRAC transient perform improvement by feedback *Datta A* + *T-AC Sep 94* 1977-1980

Nyquist envelope of interval plant family *Hollot C V* + *I-AC Feb 94* 391-396

observers for nonlin systs in steady state *Hunt L R* + *T-AC Oct 94* 2113-2118

quadratic stabilization of continuous time systs *Mahmoud M S* + *T-AC Oct 94* 2135-2139

rigid robotic manipulators robust tracking control *Man Zhihong* + *T-AC Jan 94* 154-159

slowly varying and LTI systs ident and uncertainty principles *Zames G* + *I-AC Sep 94* 1827-1838

spectral syst finite-dimens approx *Frickson M A* + *T-AC Sep 94* 1904-1909

switched-mode power converters multirate modeling and control design *Khayatian A* + *T-AC Sep 94* 1848-1852

tip mass/cable syst stabilization *Morgul O* + *T-AC Oct 94* 2140-2145

Feedback systems, cf. Output feedback, State feedback**Feedforward neural networks**

continuous-time nonlin systs adaptive control by neural nets *Fu-Chuang Chen* + *T-AC Jun 94* 1306-1310

Feedforward systems

floating platform modeling and control *Damen A A H* + *T-AC May 94* 1075-1078

Filtering

fault detect filters, eigenstructure assignment *Jaehong Park* + *T-AC Jul 94* 1521-1524

flexible joint robots global regulation approx diff *Kelly R* + *T-AC Jun 94* 1222-1224

Kalman filtering for uncertain discrete-time systs *Lihua Xie* + *T-AC Jun 94* 1310-1314

lin discrete-time systs optimum lin recursive estim *Carazo A H* + *T-AC Aug 94* 1636-1638

MIMO stochastic discrete syst state estim *Liu Danyang* + *I-AC Oct 94* 2087-2091

modified EKF *Ahmed N U* + *T-AC Jun 94* 1322-1326

multiresolutional distributed filtering *Lang Hong* *T-AC Apr 94* 853-856

optimal filtering, stochastic discrete-time systs, unknown inputs *Borisov A V* + *T-AC Dec 94* 2461-2464

stabil domains, assigned root location polynomials, convexity props *Tesi A* + *T-AC Mar 94* 668-672

stochastic adaptive prediction/MRAC *Wei Ren* + *T-AC Oct 94* 2047-2060

stochastic lin discrete systs, state estim algm *Ahmed M S* *T-AC Aug 94* 1652-1656

three-block generalized/std Riccati eqns comparison *Darouach M* + *T-AC Aug 94* 1755-1758

time-variant displacement struct and interpolation problems *Sayed A H* + *T-AC May 94* 960-976

Filtering; cf. Estimation, Nonlinear filtering**Filter noise**

modified EKF *Ahmed N U* + *T-AC Jun 94* 1322-1326

Filters; cf. Adaptive filters, Discrete-time filters**Finite automata**

discrete event systs, decentralized control, normality theorem *Lin F* + *T-AC May 94* 1089-1093

Finite wordlength effects

sampled-data systs, wordlength constraint stabil /perform *Fialho I J* + *T-AC Dec 94* 2476-2481

Flexible manufacturing systems

IMS, FCFS scheduling policy *Seidman T I* *T-AC Oct 94* 2166-2171

mfg scheduling syst regulator stabilization tech *Humes C Jr* *T-AC Jan 94* 191-196

Flexible structures

flexible joint robots with uncertain params and disturbances, tracking control *Tomei P* *T-AC May 94* 1067-1072

ident state-space freq domain approach *Bayard D S* *T-AC Sep 94* 1880-1885

multibody flexible systs elastic mode estim *Karray F* + *T-AC May 94* 1016-1020

nonlin uncertain systs tracking control *Song Y D* + *T-AC Sep 94* 1866-1871

Flight control; cf. Space vehicle control**Flow control**

zero sum Markov games *Altman F* *T-AC Apr 94* 814-818

Flow control, cf. Packet switching**Force control**

constrained robots force/motion control *Grabbe M T* + *T-AC Jan 94* 179

motion/force control robust with nonholonomic constraints *Chun Yi Su* + *T-AC Mar 94* 609-614

robotic arms, repositioning control by learning *Lucibello P* *T-AC Aug 94* 1690-1694

robot manipulators compliant force/posn regulation *Chiaverini S* + *T-AC Mar 94* 647-652

robot variable struct control schemes *Bin Yao* + *T-AC Feb 94* 371-376

stick-slip avoidance PD control *Dupont P E* *I-AC May 94* 1094-1097

tip mass/cable syst stabilization *Morgul O* + *T-AC Oct 94* 2140-2145

Forecasting, cf. Load forecasting**Formal languages**

discrete event systs, decentralized control normality theorem *Lin F* + *T-AC May 94* 1089-1093

supervisory control blocking and controllability of Petri nets *Giua A* + *T-AC Apr 94* 818-823

timed discrete-event systs supervisory control *Brandin B A* + *T-AC Feb 94* 329-342

Fourier transforms

continuous coprime factors *Treil S* *T-AC Jun 94* 1262-1263

Frequency-domain analysis

delay systs approx, Laguerre formula *Lam J* *T-AC Jul 94* 1517-1521

H₂ control syst ident least sq methods comment *Livstone M M* + *T-AC Jul 94* 1531

ident family of norms *Massoumnia M A* + *T-AC May 94* 1027-1031

ident state-space freq domain approach *Bayard D S* *T-AC Sep 94* 1880-1885

intersample props, robustness and sensitivity quantitat /qualitat anal *Feuer A* + *T-AC May 94* 1042-1047

interval lin control syst robust parametric design *Keel L H* + *T-AC Jul 94* 1524-1530

lin syst approx, time-scaling factor, Laguerre models *Wang L* + *T-AC Jul 94* 1463-1467

multirate sampled-data systs, H₂-optimal design *Qui I* + *T-AC Dec 94* 2506-2511

multivariable nonminimum phase discrete-time systs ITR procedure *Leon de la Barra S B A* *T-AC Mar 94* 574-577

nonlin feedback parametrically uncertain systs, absol stabil *Marquez H J* + *T-AC Mar 94* 664-668

Fuzzy control

book review, Handbook of Intelligent Control Neural, Fuzzy and Adaptive Approaches (White, D A, and Sofge D A, Eds, 1992) *Samad T* *T-AC Jul 94* 1534-1535

supervisory controller, fuzzy control systs *Li-Xiu Wang* *T-AC Sep 94* 1845-1847

G

Game theory

- certainty equivalence control, dyn. games. *James, M.R.*, T-AC Nov 94 2321-2324
- descriptor systs., lin. feedback closed-loop Stackelberg strategy. *Hua Xu*, +, T-AC May 94 1097-1102
- H_∞ -optimal discrete-time fixed-point and fixed-lag smoothing, game theory approach. *Theodor, Y.*, +, T-AC Sep 94 1944-1948
- H^∞ -optimal control for sing. perturbed systs., imperfect state meas. *Zigang Pan*, +, T-AC Feb 94 280-299
- lin.-quadratic zero-sum differential games for generalized state space systs. *Hua Xu*, +, T-AC Jan 94 143-147
- mixed H_2/H_∞ control, Nash game approach. *Limebeer, D.J.N.*, +, T-AC Jan 94 69-82
- noncooperative equilibria computation, relax. algms. *Uryas'ev, S.*, +, T-AC Jun 94 1263-1267
- partially obs. discrete-time nonlin. systs., risk-sensitive control and dyn. games. *James, M.R.*, +, T-AC Apr 94 780-792

Gaussian noise

- multipoint boundary value stochastic systs., stationarity/reciprocity. *Jie Chen*, +, T-AC May 94 1114-1116

Gaussian processes; cf. Gaussian noise; Linear-quadratic-Gaussian control

Geometry

- output feedback problem in lin. systs., bilinear formulation. *Syrmos, V.L.*, +, T-AC Feb 94 410-414
- uncertain nonlin. systs., struct. invariance. *Castro-Linares, R.*, +, T-AC Oct 94 2154-2158

Geometry; cf. Computational geometry; Differential geometry

Gradient methods

- path formulated optimal routing algm., time complexity. *Antonio, J.K.*, +, T-AC Sep 94 1839-1844

Graph theory

- complex discrete-time control systs., simul. automation. *Ellis, R.D.*, +, T-AC Sep 94 1795-1801
- continuous coprime factors. *Treil, S.*, T-AC Jun 94 1262-1263
- data network path formulated optimal routing, time complexity. *Antonio, J.K.*, +, T-AC Feb 94 385-391
- LP, timed marked graphs eval. *Yamada, T.*, +, T-AC Mar 94 696-698
- packet switching, flow control, rate-based, monotonicity/concavity. *Budka, K.C.*, T-AC Mar 94 544-548
- path formulated optimal routing algm., time complexity. *Antonio, J.K.*, +, T-AC Sep 94 1839-1844
- stochastic dyn. job shops/prod. planning. *Sethi, S.*, +, T-AC Oct 94 2061-2076
- stochastic timed-event graphs, superposn. props./perform. bounds. *Xiao-Lan Xie*, T-AC Jul 94 1376-1386

Graph theory; cf. Trees, graphs

H

Hermitian matrices

- polynomial J-spectral factorization. *Kwakernaak, H.*, +, T-AC Feb 94 315-328
- spectral rad. bounds in terms of Hermitian parts. *Rachid, A.*, T-AC Jan 94 196-198

Hilbert spaces

- constrained optim. in Hilbert space. *Shimizu, K.*, +, T-AC May 94 982-986
- finite-dimens. controllers designed for infinite-dimens. systs., state space. *Morris, K.A.*, T-AC Oct 94 2100-2104

 H^∞ optimization

- $2\frac{1}{2}$ -d.o.f. LQG control, rel., H^2 control. *Grimble, M.J.*, T-AC Jan 94 122-127
- book review; H_∞ Control Problem: A State Space Approach (Stoorvogel, A.; 1992). *Saberi, A.*, T-AC Jan 94 252-254
- discrete-time Riccati eqn., H_∞ control appl. *Stoorvogel, A.A.*, +, T-AC Mar 94 686-691
- discrete-time systs., mixed H_2/H_∞ control, exact convex optim. based soln. *Sznaier, M.*, T-AC Dec 94 2511-2517
- dissipative H_2/H_∞ controller synthesis. *Haddad, W.M.*, +, T-AC Apr 94 827-831
- dual-rate control syst., sampled-data, stabil. robustness. *Tongwen Chen*, T-AC Jan 94 164-167
- floating platform, modeling and control. *Damen, A.A.H.*, +, T-AC May 94 1075-1078
- H_∞ compensator design, min. order observers. *Hsu, C.S.*, +, T-AC Aug 94 1679-1681
- H_∞ control of discrete-time uncertain systs. *Geromel, J.C.*, +, T-AC May 94 1072-1075
- H_∞ control, state availability rel., pole/zero cancellations. *Miyamoto, S.*, +, T-AC Feb 94 379-381

- H_∞ multiobjective robust control, infinite-horizon. *Theodor, Y.*, +, T-AC Oct 94 2130-2134
- H_∞ -norm approx. of systs. by const. matrices. *Kavranoglu, D.*, T-AC May 94 1006-1009
- H_∞ -optim., reduced order observer based controller. *Stoorvogel, A.A.*, +, T-AC Feb 94 355-360
- H_∞ optim., time-domain constraints. *Rotstein, H.*, +, T-AC Apr 94 762-779
- H^∞ -optimal control for sing. perturbed systs., imperfect state meas. *Zigang Pan*, +, T-AC Feb 94 280-299
- LTI syst. pos. real control. *Weiglan Sun*, +, T-AC Oct 94 2034-2046
- mixed H_2/H_∞ control, Nash game approach. *Limebeer, D.J.N.*, +, T-AC Jan 94 69-82
- mixed H_2/H_∞ perform. objectives. *Kemin Zhou*, +, T-AC Aug 94 1564-1574
- mixed H_2/H_∞ perform. objectives, optimal control. *Doyle, J.*, +, T-AC Aug 94 1575-1587
- nonlin. syst. H_∞ control, output-feedback based. *Lu, W.-M.*, +, T-AC Dec 94 2517-2524
- rational transfer fn., optimal L_∞ approx. *Kavranoglu, D.*, +, T-AC Sep 94 1899-1904
- risk-averse decentralized discrete-time LEQG games. *Srikant, R.*, T-AC Apr 94 861-864
- sampled data control systs., H_∞ type problem. *Hayakawa, Y.*, +, T-AC Nov 94 2278-2284
- SISO-distributed plants, optimal mixed sensitivity. *Flamm, D.S.*, +, T-AC Jun 94 1150-1165
- state delayed systs., memoryless H^∞ controllers. *Joon Hwa Lee*, +, T-AC Jan 94 159-162
- worst case ident., H^∞ model validation. *Guoxiang Gu*, T-AC Aug 94 1657-1661
- Hurwitz stability; cf. Routh methods**
- Hysteresis nonlinearities**
- nonminimum phase first-order continuous-time systs., adaptive stabilization. *Lozano, R.*, +, T-AC Aug 94 1748-1751
- uncertain SISO min phase lin. syst., nonlin. universal servomechanism. *Ryan, E.P.*, T-AC Apr 94 753-761

I

Identification

- discrete time Hammerstein syst. ident. *Lang Zi-Qiang*, T-AC Mar 94 569-573
- signed output error adaptive ident., convergence. *Garnett, J.*, +, T-AC Jul 94 1387-1399

Identification; cf. Parameter identification, System identification

Image processing

- multiscale recursive estim., data fusion/regularization. *Chou, K.C.*, +, T-AC Mar 94 464-478

Information theory

- dyn. shock-error models, online estim. *Krishnamurthy, V.*, T-AC May 94 1129-1135

Innovations methods (stochastic processes)

- lin. discrete-time systs., optimum lin. recursive estim. *Carazo, A.H.*, +, T-AC Aug 94 1636-1638

Input-output stability

- interval plants, closed-loop hyperstability. *Foo, Y.K.*, +, T-AC Jan 94 151-154
- nonlin. discrete-time systs., input-to-state stabil. condition and global stabilization. *Kazakos, D.*, +, T-AC Oct 94 2111-2113
- robust direct adaptive controllers, normalization tech. *Gang Feng*, +, T-AC Nov 94 2330-2334
- semi-cancellable fraction transfer fns. in syst. theory. *Bourles, H.*, T-AC Oct 94 2148-2153

Integer programming

- vector discrete-event syst. controller synthesis. *Yong Li*, +, T-AC Mar 94 512-531

Integration (math.)

- delta-operator formulated discrete-time approx. *Premaratne, K.*, +, T-AC Mar 94 581-585

Integration (math.); cf. Numerical integration

Intelligent control

- book review; Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches (White, D.A., and Sofge, D.A., Eds., 1992). *Samad, T.*, T-AC Jul 94 1534-1535
- continuous-time nonlin. systs. adaptive control by neural nets. *Fu-Chun Chen*, +, T-AC Jun 94 1306-1310
- rule-based incremental control, machine learning. *Luzeaux, D.*, T-AC Jan 94 1166-1171

Interconnected systems

- discrete-time systs., stabil. robustness anal. *Karan, M.*, +, T-AC May 94 991-995

- interconnected systs, decentralized adaptive regulation *Changyun Wen* *T-AC Oct 94* 2163-2166
- large scale interconnected systs with delays, decentralized stabilization *Zhongzhi Hu* *T-AC Jan 94* 180-182
- nonlin discrete-time systs, input-to-state stabil condition and global stabilization *Kazakos, D., +*, *T-AC Oct 94* 2111-2113
- uncertain nonlin interconnected systs, time-varying state delay, stabilizing control *Mahmoud MS* *T-AC Dec 94* 2484-2488
- interconnected systems; cf. Cascade systems, Large-scale systems
- interference; cf. Noise
- interpolation
- four-block problem, recursive Schur-based soln *Constantinescu T +* *T-AC Jul 94* 1476-1481
- polynomial J-spectral factorization *Kwakernaak H +* *T-AC Feb 94* 315-328
- SISO systs, robust stabilization *Olbrot A W +* *T-AC Mar 94* 652-657
- time-variant displacement struct and interpolation problems *Sayed A H +* *T-AC May 94* 960-976
- Interpolation; cf. Spline functions
- Inverse problems
- discrete-time q-Markov cover generation inverse Lyapunov method *Sreeram V +* *T-AC Feb 94* 381-385
- lin discrete syst pole assignment *Benzaouia A* *T-AC Oct 94* 2091-2095
- Inverse problems, linear; cf. Matrix inversion
- Irreducible realizations, cf. Minimal realizations
- Iterative methods
- data network path formulated optimal routing, time complexity *Antonio J K +* *T-AC Feb 94* 385-391
- dyn syst simul, fast parallel recursive aggregation *Tsai W K +* *T-AC Mar 94* 534-540
- fault detect filters eigenstructure assignment *Jaehong Park +* *T-AC Jul 94* 1521-1524
- interval lin control syst robust parametric design *Keel I II +* *T-AC Jul 94* 1524-1530
- least absol values estim, computational aspects *Fiodorov E D* *T-AC Mar 94* 626-630
- matrix sign ln calc iterated fraction expansion *Koc C K +* *T-AC Aug 94* 1644-1647
- path formulated optimal routing algm time complexity *Antonio J K +* *T-AC Sep 94* 1839-1844
- pole placement by output feedback *Iee T H +* *T-AC Mar 94* 565-568
- queuing networks anal projection techs *Mooi Choo Chuah* *T-AC Aug 94* 1588-1599
- Riccati eqn discrete alg iter matrix bounds *Komaroff N* *T-AC Aug 94* 1676-1678
- state-constrained optimal control generalised dual quasi-Newton algm *Shimizu K +* *T-AC May 94* 982-986
- textured iter algm for tridiagonal lin eqns *Tian-Shen Tang* *T-AC Mar 94* 592-596

J

Jump parameter systems

- discrete-time Markovian jump lin systs, lin min MSF estim *Costa O I V* *T-AC Aug 94* 1685-1689
- jump lin systs coupled Riccati eqns *Abou Kandil H +* *T-AC Aug 94* 1631-1636
- jump-Markov systs stabilizing control law *Dufour F +* *T-AC Nov 94* 2354-2357

K

Kalman filtering

- autonomous vision-based mobile robot *Baumgartner E T +* *T-AC Mar 94* 493-502
- lyn shock-error models, online estim *Krishnamurthy V* *T-AC May 94* 1129-1135
- uaranteed estim problem, Kalman-Bucy filter appl *Golovan A +* *T-AC Jun 94* 1282-1286
- uaranteed param estim problem with uncertain stats, Kalman-Bucy filter accuracy *Matasov A I* *T-AC Mar 94* 635-639
- in sing perturbed systs, Kalman filtering *Gajic Z +* *T-AC Sep 94* 1952-1955
- IMO stochastic discrete syst state estim *Liu Danyang +* *T-AC Oct 94* 2087-2091
- modified EKF *Ahmed N U +* *T-AC Jun 94* 1322-1326
- ultiscale recursive estim, data fusion/regularization *Chou K C +* *T-AC Mar 94* 464-478
- ultiscale systs, Kalman filters, Riccati eqns *Chou K C +*, *T-AC Mar 94* 479-492

- optimal cost control/filtering, uncertain lin systs *Petersen, I R., +* *T-AC Sep 94* 1971-1977
- three-block generalized/std Riccati eqns comparison *Darouach, M., +*, *T-AC Aug 94* 1755-1758
- uncertain discrete time systs appl *Lihua Xie, +* *T-AC Jun 94* 1310-1314
- Kalman filtering; cf. Adaptive Kalman filtering
- Knowledge-based systems
- rule-based incremental control, machine learning *Luzeaux D, T-AC Jun 94* 1166-1171

L

Laguerre processes

- lin syst approx, time-scaling factor, Laguerre models *Wang I +* *T-AC Jul 94* 1463-1467

Land navigation

- autonomous vision-based mobile robot *Baumgartner E T +* *T-AC Mar 94* 493-502

Laplace transforms

- transfer fn param ident in freq domain *Pintelon R +* *T-AC Nov 94* 2245-2260

Large-scale integration; cf. Very-large-scale integration

Large-scale systems

- complex discrete-time control systs, simul automation *Ellis R D +* *T-AC Sep 94* 1795-1801
- H[∞]-optimal control for sing perturbed systs, imperfect state meas *Zigang Pan +* *T-AC Feb 94* 280-299
- large scale interconnected systs with delays, decentralized stabilization *Zhongzhi Hu* *T-AC Jan 94* 180-182
- sing perturbed control systs and nonlin differential-alg eqns *Krishnan H +* *T-AC May 94* 1079-1084

Large-scale systems; cf. Interconnected systems, Reduced-order systems

Learning control systems

- P-type learning control *Saab S S* *T-AC Nov 94* 2298-2302
- rule-based incremental control, machine learning *Luzeaux D* *T-AC Jun 94* 1166-1171

Learning systems

- autonomous vision-based mobile robot *Baumgartner E T +* *T-AC Mar 94* 493-502
- continuous-time nonlin systs adaptive control by neural nets *Fu-Chuang Chen +* *T-AC Jun 94* 1306-1310
- I/O map inverse pass learning update-spline-smoothing *Heiss M* *T-AC Feb 94* 259-268
- robotic arms, repositioning control by learning *Lucibello P* *T-AC Aug 94* 1690-1694
- rule-based incremental control, machine learning *Luzeaux D* *T-AC Jun 94* 1166-1171

Least-mean-square methods

- H_∞ control syst ident least sq methods comment *Livstone M M +* *T-AC Jul 94* 1531

Least-squares methods

- adaptive pole placement without excitation probing sigs *Lozano R +* *T-AC Jan 94* 47-58
- bilinear systs, adaptive stabilization, least sq *Xi Sun +* *T-AC Jan 94* 207-211
- convergence of least sq estim in white noise *Nassiri-Toussi K +* *T-AC Feb 94* 364-368
- discrete-time adaptive nonlin syst least sq estimator *Kanellakopoulos I* *T-AC Nov 94* 2362-2365
- efficient algm for least sq estim *Rafajlowicz E +* *T-AC Jun 94* 1241-1243
- extended Chandrasekhar recursions *Sayed A H +* *T-AC Mar 94* 619-623
- first-order nonlin syst adaptive control *Brogliato B +* *T-AC Aug 94* 1764-1768
- guaranteed estim problem, Kalman-Bucy filter appl *Golovan A, +* *T-AC Jun 94* 1282-1286
- least absol values estim, computational aspects *Fiodorov E D* *T-AC Mar 94* 626-630
- multibody flexible systs, elastic mode estim *Karray F +* *T-AC May 94* 1016-1020
- nonminimum phase first-order continuous-time systs, adaptive stabilization *Lozano R +* *T-AC Aug 94* 1748-1751
- sing-free multivariable MRAC *Moctezuma R G +* *T-AC Sep 94* 1856-1860
- stochastic adaptive prediction/MRAC *Wei Ren +* *T-AC Oct 94* 2047-2060
- time-varying params, bounded error ident *Bittanti S +* *T-AC May 94* 1106-1110

Lie algebras

- rigid spacecraft systs controllability, central gravitational field *Lian K-Y +* *T-AC Dec 94* 2426-2441

- robot kinematics, computational aspects of product-of-exponentials formula. *Park, F.C.*, *T-AC Mar 94* 643-647
- Linear algebra; cf. Matrices**
- Linear approximation**
approx. feedback lin., stabil. *Kwanghee Nam, +*, *T-AC Nov 94* 2311-2314
brushless DC motors, feedback-lin control charactn. *In-Joong Ha, +*, *T-AC Mar 94* 673-677
feedback lin. systs. adaptive tracking. *Martino, R., +*, *T-AC Jun 94* 1314-1319
input-output lin., state equivalence and decoupling. *In-Joong Ha, +*, *T-AC Nov 94* 2269-2274
MIMO nonlin. syst. I/O pseudolinarization. *Lawrence, D.A., +*, *T-AC Nov 94* 2207-2218
parametrically uncertain nonlin. systs., robust stabilisation. *Schoenwald, D.A., +*, *T-AC Aug 94* 1751-1755
stochastic dyn. systs., exact lin. *Socha, L.*, *T-AC Sep 94* 1980-1984
uncertain nonlin. systs., struct. invariance. *Castro-Linares, R., +*, *T-AC Oct 94* 2154-2158
- Linear programming**
closed Jackson queueing network decentralized control. *Dye-Jyun Ma, +*, *T-AC Jul 94* 1460-1463
model validation, time-domain approach. *Poolla, K., +*, *T-AC May 94* 951-959
prod. control methods, distributed, stabil. and perform. *Sharifnia, A.*, *T-AC Apr 94* 725-737
queueing networks and scheduling policies, perform. bounds. *T-AC Aug 94* 1600-1611
queueing networks anal., projection techs. *Mooi Choo Chuah, T-AC Aug 94* 1588-1599
timed marked graphs, perform. eval. *Yamada, T., +*, *T-AC Mar 94* 696-698
vector discrete-event syst. controller synthesis. *Yong Li, +*, *T-AC Mar 94* 512-531
- Linear programming; cf. Integer programming**
- Linear-quadratic control**
biased/unbiased controllers LTR design. *Turan, L., +*, *T-AC Mar 94* 601-605
descriptor systs., lin. feedback closed-loop Stackelberg strategy. *Hua Xu, +*, *T-AC May 94* 1097-1102
exponential lin. quadratic optimal control, discounting. *Hopkins, W.E., Jr.*, *T-AC Jan 94* 175-178
infinite horizon LQ control, soln approx. *Schochetman, I.E., +*, *T-AC Mar 94* 596-601
infinite time-varying LQ-problem, approx soln. *Engwerda, J.C.*, *T-AC Jan 94* 235-238
lin. quadratic control, convex prog., num. method. *Peres, P.L.D., +*, *T-AC Jan 94* 198-202
LQ optimal regulators, stabil. robustness. *Dohyoung Chung, +*, *T-AC Aug 94* 1698-1702
optimal cost control/filtering, uncertain lin systs. *Petersen, I.R., +*, *T-AC Sep 94* 1971-1977
Riccati eqn., discrete-time alg., closed loop eigenvalues. *Lin-Zhang Lu, +*, *T-AC Aug 94* 1682-1685
robust lin. quadratic designs, real param. uncertainty. *Douglas, J., +*, *T-AC Jan 94* 107-111
- Linear-quadratic-Gaussian control**
 $2\frac{1}{2}$ -d.o.f. LQG control, rel. H^2 control. *Grimble, M.J.*, *T-AC Jan 94* 122-127
asynchronous sampled data syst control, design. *Voulgaris, P.*, *T-AC Jul 94* 1451-1455
discrete-time compensators, LTR. *Tadjine, M., +*, *T-AC Jun 94* 1259-1262
infinite horizon optimal control of stochastic systs. *Runolfsson, T.*, *T-AC Aug 94* 1551-1563
LQG controllers, stable, weighting and covariance matrices. *Halevi, Y.*, *T-AC Oct 94* 2104-2106
risk-averse decentralized discrete-time LEQG games. *Srikant, R.*, *T-AC Apr 94* 861-864
- Linear systems**
adaptive stochastic control algms., unmodified, robustness. *Radenkovic, M.S., +*, *T-AC Feb 94* 396-400
block multirate input-output model for sampled-data control systs. *Jakubowski, A.M., +*, *T-AC May 94* 1145-1147
book review; Controlled and Conditioned Invariants in Linear System Theory (Basile, G., and Marro, G.; 1992). *Schumacher, J.M.*, *T-AC Jan 94* 250-251
book review; Linear Multivariable Control: Algebraic Analysis and Synthesis Methods (Vardulakis, A.I.G.; 1991). *Lewis, F.L.*, *T-AC Jul 94* 1536
book review; Linear System Theory (Rugh, W.J.; 1993). *Khalil, H.K.*, *T-AC Dec 94* 2528-2529
book review; New Tools for Robustness of Linear Systems (Barmish, B.R.; 1994). *T-AC Dec 94* 2525-2526
CARMA plants, stabilizing I-O receding horizon control. *Chisci, L., +*, *T-AC Mar 94* 614-618
certainty equivalence control, dyn. games. *James, M.R.*, *T-AC Nov 94* 2321-2324
comments on "Stability margin evaluation for uncertain linear systems (by C. Gong and S. Thompson, Jun 94 548-550). *Su, J.-H.*, *T-AC Dec 94* 2523-2524
comments on "System zeros determination from an unreduced matrix fraction description" (by K.S. Yeung and C.-M. Kwan, Nov 94 1695-1697). *Ferreira, P.M.G.*, *T-AC Nov 94* 2367
comments, with reply, on "Vector norms as Lyapunov functions for linear systems" (by H. Kiendl et al., Jun 92 839-842). *Hamed, A.*, *T-AC Dec 94* 2522-2523
common Lyapunov fn., stable LTI systs., commuting A-matrices. *Narendra, K.S., +*, *T-AC Dec 94* 2469-2471
constrained continuous-time systs. with cone props., pos. invariant sets. *Tarbouriech, S., +*, *T-AC Feb 94* 401-405
continuous coprime factors. *Treil, S.*, *T-AC Jun 94* 1262-1263
continuous time nonminimum phase lin. systs., gain margin improvement. *Wei-Yong Yan, +*, *T-AC Nov 94* 2347-2354
continuous-time systs., global stabilization, bounded controls. *Sussmann, H.J., +*, *T-AC Dec 94* 2411-2425
corrections to "Positively invariant sets for constrained continuous-time systems with cone properties" (Feb 94 401-405). *Tarbouriech, S.*, *T-AC Aug 94* 1771
DD discrete time systs., pole assignment. *Benzaouia, A.*, *T-AC Oct 94* 2091-2095
decentralized stable factors and parameterization of decentralized controllers. *Dale, R., +*, *T-AC Feb 94* 347-351
diagonal decoupling, dyn. output feedback/const. precompensator. *Eldem, V.*, *T-AC Mar 94* 503-511
discrete-time lin. systs., pos. invariant sets. *De Santis, E.*, *T-AC Jan 94* 245-249
discrete-time Markovian jump lin. systs., lin. min. MSE estim. *Costa, O.L.V.*, *T-AC Aug 94* 1685-1689
discrete-time q-Markov cover models. *Sreeram, V., +*, *T-AC May 94* 1102-1105
discrete-time systs., optimum lin. recursive estim. *Carazo, A.H., +*, *T-AC Aug 94* 1636-1638
discrete-time systs., perturb., comment. *Eslami, M.*, *T-AC Aug 94* 1768-1769
discrete-time uncertain systs., ultimate boundedness control, Lyapunov. *Blanchini, F.*, *T-AC Feb 94* 428-433
distributed delay lin. systs., feedback stabilisation. *Feng Zheng, +*, *T-AC Aug 94* 1714-1718
D-stabil., robust, for lin. uncertain discrete delay systs. *Te-Jen Su, +*, *T-AC Feb 94* 425-428
dyn. syst. simul., fast parallel recursive aggregation. *Tsai, W.K., +*, *T-AC Mar 94* 534-540
finite-dimens. model validation, output error, test horizon. *Hoai Nghu Duong, +*, *T-AC Jan 94* 102-106
full-order/low-order observers, LTI plant, unified singular syst. based theory. *Cobb, J.D.*, *T-AC Dec 94* 2497-2502
 H^∞ -optimal control for sing. perturbed systs., imperfect state meas. *Zigang Pan, +*, *T-AC Feb 94* 280-299
impulse differential lin. systs., constrained controllability. *Benzaid, Z., +*, *T-AC May 94* 1064-1066
input-output lin., state equivalence and decoupling. *In-Joong Ha, +*, *T-AC Nov 94* 2269-2274
interval LTI systs., stabilization. *Kehui Wei, T-AC Jan 94* 22-32
jump lin. systs., coupled Riccati eqns. *Abou-Kandil, H., +*, *T-AC Aug 94* 1631-1636
lin. min-phase plants, stabilizing controllers parameterization. *Glaria J.J., +*, *T-AC Feb 94* 433-434
LTI discrete-time systs., robust stabil. anal. *Karan, M., +*, *T-AC Mar 94* 991-995
min. lin. plants parameterization. *Davis, L.D., +*, *T-AC Apr 94* 849-852
multivariable systs., fund. notion of equivalence. *Pugh, A.C., +*, *T-AC May 94* 1141-1145
nonlin. design of adaptive controllers. *Krstic, M., +*, *T-AC Apr 94* 738-752
output feedback problem in lin. systs., bilinear formulation. *Syrmos, L., +*, *T-AC Feb 94* 410-414
param. hypothesis testing, stochastic lin. systs., stat. sampling. *Duan, T.E., +*, *T-AC Jan 94* 118-122
perturbed lin. systs., upper covariance bounds. *Bolzern, P., +*, *T-AC Mar 94* 623-626
pos. of dyn. systs., nonpositive coeff. matrices. *d'Alessandro, P., +*, *T-AC Jan 94* 131-134
P-type learning control. *Saab, S.S.*, *T-AC Nov 94* 2298-2302
SISO LTI discrete-time systs., param.-adaptive controller. *Kretsselmeyer, G.*, *T-AC Sep 94* 1819-1826

slowly varying and LTI sys t s, ident and uncertainty principles *Zames G* + *T-AC Sep 94* 1827-1838

stabil of uncertain lin sys t s, saturating actuators *Jin-Hoon Kum* + *T-AC Jan 94* 202-207

state delayed sys t s, memoryless H^∞ controllers *Joon Hwa Lee* + *T-AC Jan 94* 159-162

stochastic control, lin -exponential-Gaussian *Chih-Hai Fan* + *T-AC Oct 94* 1986-2003

switching surface design for multivariable VSS *Ju-Jang Lee* + *T-AC Feb 94* 414-419

time-invariant lin sys t s, robust controller synthesis *Rantzer A* + *T-AC Sep 94* 1802-1808

time-invariant sys t s, pos real control *Weigian Sun* + *T-AC Oct 94* 2034-2046

uncertain lin sys t s, delay depend, robust stabil *Bugong Xu* *T-AC Nov 94* 2365

uncertain lin sys t s multivariable stabil margin eval *Gong C* + *T-AC Mar 94* 548-550

uncertain syst robust stabil anal, nonquadratic Lyapunov fns *Zelentsovsky A L* *T-AC Jan 94* 135-138

uncertain syst robust stabil, guardian map approach *Shuoh Rern* + *T-AC Jan 94* 162-164

unknown input lin sys t s, full-order observer design *Darouach M* + *T-AC Mar 94* 606-609

unknown input observers design *Darouach M* *T-AC Mar 94* 698-699

Wiener sys t s nonparametric ident *Greblicki W* *T-AC Oct 94* 2077-2086

worst-case ident anal, BIBO robustness *Partington J R* + *T-AC Oct 94* 2171-2176

LMS, cf. Least-mean-square methods

Load forecasting

alternating renewal elec load models *El-Ferik S* + *T-AC Jun 94* 1184-1196

Load frequency control; cf. Power generation control load frequency

Load modeling

alternating renewal elec load models *El-Ferik S* + *T-AC Jun 94* 1184-1196

LQ control; cf. Linear-quadratic control

LQG control; cf. Linear-quadratic-Gaussian control

Lyapunov matrix equations

discrete-time q-Markov cover generation, inverse Lyapunov method *Sreeram V* + *T-AC Feb 94* 381-385

large sparse Lyapunov eqns approx soln *Gudmundsson I* + *T-AC May 94* 1110-1114

multivariable sys t s quantitat robustness measures *Hornig-Giou Chen* + *T-AC Apr 94* 807-810

sing perturbed syst block-diagonalization Taylor expansion *Derbel N* + *T-AC Jul 94* 1429-1431

Lyapunov methods

asynchronous sys t s with Poisson transits stabil *Leland R P* *T-AC Jan 94* 182-185

attitude control problem comment *Fjellstad O-E* + *T-AC Mar 94* 699-700

book review Applications of Lyapunov Methods in Stability (Halanay A, and Rasvan V 1993) *Popov V M* *T-AC Dec 94* 2526-2527

comments with reply on 'Vector norms as Lyapunov functions for linear systems' (by H Kiendl et al Jun 92 839-842) *Hmamed A* *T-AC Dec 94* 2522-2523

common Lyapunov fn stable LTI sys t s, commuting A-matrices *Narendra K S* + *T-AC Dec 94* 2469-2471

corrections to An improved Razumikhin-type theorem and its application" (Apr 94 839-841) *Xu B* *T-AC Nov 94* 2368

corrections to "Lyapunov stability of a class of discrete-event systems" (Feb 94 269-279) *Passino K M* + *T-AC Jul 94* 1531

discrete-time adaptive nonlin syst least sq estimator *Kanellakopoulos I* *T-AC Nov 94* 2362-2365

discrete-time syst exponential stabil, observer design *Aitken V C* + *T-AC Sep 94* 1959-1962

discrete-time sys t s, generalized q-Markov cover models *Sreeram V* *T-AC Dec 94* 2502-2505

discrete-time sys t s perturb comment *Eslami M* *T-AC Aug 94* 1768-1769

discrete-time uncertain sys t s, ultimate boundedness control, Lyapunov *Blanchini F* *T-AC Feb 94* 428-433

finite-time optimal control of bilinear sys t s, successive approx procedure *Aganovic Z* + *T-AC Sep 94* 1932-1935

interval matrices, necessary and sufficient conditions for Hurwitz/Schur stabil *Kaining Wang* + *T-AC Jun 94* 1251-1255

logical discrete event sys t s, Lyapunov stabil *Passino K M* + *T-AC Feb 94* 269-279

LOG controllers, stable, weighting and covariance matrices *Halevi Y* *T-AC Oct 94* 2104-2106

manipulators with parametric uncertainty, robust control *Keun-Mo Koo* + *T-AC Jun 94* 1230-1233

multivariable sys t s, quantitat robustness measures *Hornig-Giou Chen* + *T-AC Apr 94* 807-810

nonlin design of adaptive controllers *Krstic M* + *T-AC Apr 94* 738-752

nonnegative feedback control, oscills stabilisation *Zaslavsky B* *T-AC Jun 94* 1273-1276

output feedback stabilization of nonlin sys t s *Tsinias J* + *T-AC Apr 94* 806

parab sys t s direct adaptive control *Keum Shik Hong* + *T-AC Oct 94* 2018-2033

parametrically uncertain nonlin sys t s, robust stabilisation *Schoenwald, D A*, + *T-AC Aug 94* 1751-1755

perturbed lin sys t s, upper covariance bounds *Bolzern P* + *T-AC Mar 94* 623-626

Razumikhin-type theorem improved, appls *Bugong Xu* + *T-AC Apr 94* 839-841

robot controller adaptive, Lyapunov stabil *Egeland O* + *T-AC Aug 94* 1671-1673

robot manipulators, compliant, force/posn regulation *Chiaverini S* + *T-AC Mar 94* 647-652

robust strong stabilization via modified Popov controller synthesis *Wang Y W* + *T-AC Nov 94* 2284-2287

second-order dyn sys t s, dissipative controller designs *Morris K A* + *T-AC May 94* 1056-1063

stabil theory of nonsmooth sys t s *Shevitz D* + *T-AC Sep 94* 1910-1914

structurally uncertain sys t s optimal Lyapunov fns *Olas A* *T-AC Jan 94* 167-171

uncertain nonlin interconnected sys t s, time-varying state delay, stabilizing control *Mahmoud M S* *T-AC Dec 94* 2484-2488

uncertain syst robust stabil anal, nonquadratic Lyapunov fns *Zelentsovsky A I* *T-AC Jan 94* 135-138

uncertain sys t s, asymptotic stabil region estim *Han Ilo Choi* + *T-AC Nov 94* 2275-2278

M

Machine vision; cf. Robots, vision systems

Manipulators

book review Robot Dynamics and Control" (Spong, M W, and Vidyasagar, M, 1989) *Pota H R* + *T-AC Aug 94* 1774-1776

constrained robots force/motion control *Grabbe M T* + *T-AC Jan 94* 179

contact stabil of simple posn controllers effect of time delay and discrete control *Fiala J* + *T-AC Apr 94* 870-873

deterministic nonlin syst ident *Shir-Kuan Lin* *I-AC Sep 94* 1886-1893

Manipulators, dynamics

compliant manipulators stable force/posn regulation *Chiaverini S* + *T-AC Mar 94* 647-652

flexible joint robots with uncertain params and disturbances, tracking control *Tomei P* *T-AC May 94* 1067-1072

globally convergent adaptive controller *Mahmoud M S* + *T-AC Jan 94* 148-151

mobile manipulator locomotion/manipulation coord *Yamamoto Y* + *T-AC Jun 94* 1326-1332

rigid robotic manipulators, robust tracking control *Man Zhihong* + *T-AC Jan 94* 154-159

robot variable struct control schemes *Bin Yao* + *T-AC Feb 94* 371-376

robust MIMO terminal sliding-mode control, rigid robotic manipulators *Zhihong M* + *T-AC Dec 94* 2464-2469

single-arm dyn, robust variable struct /and switching- Σ adaptive control *Li-Wen Chen* + *T-AC Aug 94* 1621-1626

Manipulators, kinematics

flexible joint robots with uncertain params and disturbances, tracking control *Tomei P* *T-AC May 94* 1067-1072

kinematics, product-of-exponentials formula computational aspects *Park F C* *T-AC Mar 94* 643-647

Manipulators, motion planning

mobile manipulator locomotion/manipulation coord *Yamamoto Y* + *T-AC Jun 94* 1326-1332

Manufacturing planning

stochastic dyn job shops/prod planning *Sethi S* + *T-AC Oct 94* 2061-2076

Manufacturing scheduling

distributed prod control methods *Sharifnia, A* *T-AC Apr 94* 725-737

failure-prone mfg sys t s, structural props of optimal prod controllers *Jian-Qiang Hu* + *T-AC Mar 94* 640-643

FMS, FCFS scheduling policy *Seidman T I* *T-AC Oct 94* 2166-2171

mfg scheduling syst, regulator stabilization tech *Humes C Jr* *T-AC Jan 94* 191-196

mfg sys t s with prod rate-depend failure rates, prod control *Liberopoulos G* + *T-AC Apr 94* 889-895

- mfg. syst. with unreliable machines, stabil. anal. *Han-Fu Chen*, +, *T-AC Mar 94* 681-686
- queueing networks and scheduling policies, perform. bounds. *T-AC Aug 94* 1600-1611
- Manufacturing scheduling; cf. Production control**
- Markov processes**
- 2D stochastic approx., convergence/diff. eqn. limit. *Dye-Jyun Ma*, +, *T-AC Jul 94* 1439-1442
- alternating renewal elec. load models. *El-Ferik, S.*, +, *T-AC Jun 94* 1184-1196
- continuous-time multivariable syst. reduction. *Krajewski, W.*, +, *T-AC Oct 94* 2126-2129
- deterministic/stochastic queueing networks anal. *Cheng-Shang Chang*, *T-AC May 94* 913-931
- discrete-time Markovian jump lin. systs., lin. min. MSE estim. *Costa, O.L.V.*, *T-AC Aug 94* 1685-1689
- discrete-time q-Markov cover generation, inverse Lyapunov method. *Sreeram, V.*, +, *T-AC Feb 94* 381-385
- discrete-time q-Markov cover models. *Sreeram, V.*, +, *T-AC May 94* 1102-1105
- discrete-time systs., generalized q-Markov cover models. *Sreeram, V.*, *T-AC Dec 94* 2502-2505
- flow control, zero sum Markov games. *Altman, E.*, *T-AC Apr 94* 814-818
- jump lin. systs., coupled Riccati eqns. *Abou-Kandil, H.*, +, *T-AC Aug 94* 1631-1636
- jump-Markov systs., stabilizing control law. *Dufour, F.*, +, *T-AC Nov 94* 2354-2357
- mfg. systs. with prod. rate-depend. failure rates, prod. control. *Liberopoulos, G.*, +, *T-AC Apr 94* 889-895
- queueing networks and scheduling policies, perform. bounds. *T-AC Aug 94* 1600-1611
- Mathematical programming; cf. Dynamic programming; Linear programming**
- Mathematics; cf. Algebra**
- Matrices**
- bilinear systs., state space anal., orthogonal series approach. *Paraskevopoulos, P.N.*, +, *T-AC Apr 94* 793-797
- book review; Linear Multivariable Control: Algebraic Analysis and Synthesis Methods (Vardulakis, A.I.G., 1991) *Lewis, F.L.*, *T-AC Jul 94* 1536
- comments on "Stabilization via static output feedback". *Pimpalkhare, A.A.*, +, *T-AC May 94* 1148
- comments on "System zeros determination from an unreduced matrix fraction description" (by K.S. Yeung and C.-M. Kwan, Nov 93 1695-1697). *Ferreira, P.M.G.*, *T-AC Nov 94* 2367
- common Lyapunov fn., stable LTI systs., commuting A-matrices *Narendra, K.S.*, +, *T-AC Dec 94* 2469-2471
- constrained continuous-time systs with cone props., pos. invariant sets. *Tarbouriech, S.*, +, *T-AC Feb 94* 401-405
- continuous alg. Riccati eqn. eigenvalues. *Komaroff, N.*, *T-AC Mar 94* 532-534
- corrections to "Positively invariant sets for constrained continuous-time systems with cone properties" (Feb 94 401-405). *Tarbouriech, S.*, *T-AC Aug 94* 1771
- descriptor systs control by output feedback. *Lovass-Nagy, V.*, +, *T-AC Jul 94* 1507-1509
- diagonal decoupling, dyn. output feedback/const. precompensator. *Eldem, V.*, *T-AC Mar 94* 503-511
- discrete-time lin. systs., pos. invariant sets. *De Santis, E.*, *T-AC Jan 94* 245-249
- discrete-time Riccati eqn., H_∞ control appl. *Stoorvogel, A.A.*, +, *T-AC Mar 94* 686-691
- discrete-time systs., perturb., comment *Eslami, M.*, *T-AC Aug 94* 1768-1769
- distributed algms., random proc. failures. *Papavassilopoulos, G.P.*, *T-AC May 94* 1032-1036
- distributed delay lin. systs., feedback stabilisation. *Feng Zheng*, +, *T-AC Aug 94* 1714-1718
- disturbance decoupled observer design. *Hou, M.*, +, *T-AC Jun 94* 1338-1341
- disturbance decoupling, constrained Sylvester eqns. *Syrmos, V.L.*, *T-AC Apr 94* 797-803
- dyn. syst. simul., fast parallel recursive aggregation. *Tsal, W.K.*, +, *T-AC Mar 94* 534-540
- eigenstructure assignment by decentralized output feedback. *Guang-Ren Duan*, *T-AC May 94* 1009-1014
- extended Chandrasekhar recursions. *Sayed, A.H.*, +, *T-AC Mar 94* 619-623
- extended horizon liftings for nonminimum-phase systs. *Bayard, D.S.*, *T-AC Jun 94* 1333-1338
- fault detect. filters, eigenstructure assignment. *Jaehong Park*, +, *T-AC Jul 94* 1521-1524
- four-block problem, recursive Schur-based soln. *Constantinescu, T.*, +, *T-AC Jul 94* 1476-1481
- H_∞ control of systs. under norm bounded uncertainties. *Gu, K.*, *T-AC Jun 94* 1320-1322
- interval LTI systs., stabilization. *Kehui Wei*, *T-AC Jan 94* 22-32
- interval matrices, necessary and sufficient conditions for Hurwitz/Schur stabil. *Kaining Wang*, +, *T-AC Jun 94* 1251-1255
- interval matrices stabil. conditions. *Sezer, M.E.*, +, *T-AC Feb 94* 368-371
- jump lin. systs., coupled Riccati eqns. *Abou-Kandil, H.*, +, *T-AC Aug 94* 1631-1636
- Kalman filtering for uncertain discrete-time systs. *Lihua Xie*, +, *T-AC Jun 94* 1310-1314
- lin. discrete syst. pole assignment. *Benzaouia, A.*, *T-AC Oct 94* 2091-2095
- MIMO nonlin. syst. I/O pseudolinearization. *Lawrence, D.A.*, +, *T-AC Nov 94* 2207-2218
- min. lin. plants parameterization. *Davis, L.D.*, +, *T-AC Apr 94* 849-852
- min. stochastic realizations, parameterization. *Ferrante, A.*, *T-AC Oct 94* 2122-2126
- mobile manipulator locomotion/manipulation coord. *Yamamoto, Y.*, +, *T-AC Jun 94* 1326-1332
- multipoint boundary value stochastic systs., stationarity/reciprocity *Jie Chen*, +, *T-AC May 94* 1114-1116
- multiscale systs., Kalman filters, Riccati eqns. *Chou, K.C.*, +, *T-AC Mar 94* 479-492
- nonlin. systs, vibr. control by AP-forcing *Balestrino, A.*, +, *T-AC Jun 94* 1255-1258
- periodic discrete-time Riccati eqn., num. soln. *Hench, J.J.*, +, *T-AC Jun 94* 1197-1210
- pole placement by output feedback *Lee, T.H.*, +, *T-AC Mar 94* 565-568
- pos. of dyn. systs., nonpositive coeff. matrices. *d'Alessandro, P.*, +, *T-AC Jan 94* 131-134
- rational transfer fn., optimal L_∞ approx. *Kavranoglu, D.*, +, *T-AC Sep 94* 1899-1904
- Riccati eqn., discrete alg., iter. matrix bounds *Komaroff, N.*, *T-AC Aug 94* 1676-1678
- Riccati eqn., discrete-time alg., closed loop eigenvalues *Lin-Zhang Lu*, +, *T-AC Aug 94* 1682-1685
- robot controller, adaptive, Lyapunov stabil *Egeland, O.*, +, *T-AC Aug 94* 1671-1673
- robust nonsingularity problem, SSV *Ghwan-Lu Tseng*, +, *T-AC Oct 94* 2118-2122
- sign fn. calc., iterated fraction expansion *Koc, C.K.*, +, *T-AC Aug 94* 1644-1647
- sing.-free multivariable MRAC *Moctezuma, R.G.*, +, *T-AC Sep 94* 1856-1860
- sing. root distrib. problems, displacement struct. approach *Pal, D.*, +, *T-AC Jan 94* 238-245
- state deadbeat control problem, general soln. *Eldem, V.*, +, *T-AC May 94* 1002-1006
- textured iter. algms. for tridiagonal lin. eqns. *Tian-Shen Tang*, *T-AC Mar 94* 592-596
- time-invariant lin. systs. controllability/observability *Kaining Wang*, +, *T-AC Jul 94* 1443-1447
- time-variant displacement struct. and interpolation problems. *Sayed, A.H.*, +, *T-AC May 94* 960-976
- tridiagonal symmetric interval matrices, eigenvalues. *Commercon, J.C.*, *T-AC Feb 94* 377-379
- uncertain systs., sampled-data controller design. *Dolphus, R.M.*, *T-AC May 94* 1036-1042
- unknown input observers design *Darouach, M.*, *T-AC Mar 94* 698-699
- Matrices; cf. Covariance matrices; Hermitian matrices; Lyapunov matrix equations; Polynomial matrices; Toeplitz matrices; Transfer function matrices**
- Matrix decomposition/factorization**
- constrained Riccati eqn. factorizations. *Weiss, M.*, *T-AC Mar 94* 677-681
- descriptor systs. regularization by output feedback. *Bunse-Gerstner, A.*, +, *T-AC Aug 94* 1742-1748
- H_∞ -norm approx. of systs. by const. matrices *Kavranoglu, D.*, *T-AC Nov 94* 1006-1009
- least sq. estim., efficient algm. *Rafajlowicz, E.*, +, *T-AC Jun 94* 1241-1243
- polynomial J-spectral factorization *Kwakernaak, H.*, +, *T-AC Feb 94* 315-328
- Matrix inversion**
- structurally uncertain systs., optimal Lyapunov fns. *Olas, A.*, *T-AC Jan 94* 167-171
- Matrix multiplication**
- trace of matrix product, inequalities. *Fang, Y.*, +, *T-AC Dec 94* 2489-2490
- Maximum-likelihood estimation**
- dyn. shock-error models, online estim. *Krishnamurthy, V.*, *T-AC Mar 94* 1129-1135

Mean-square-error methods

discrete-time Markovian jump lin systs, lin min MSE estim *Costa O L V*, *T-AC Aug 94* 1685-1689

Mechanical factors

stick-slip avoidance, PD control *Dupont P E*, *T-AC May 94* 1094-1097

Mechanical factors; cf. Biomechanics**Mechanical systems**

motion/force control, robust, with nonholonomic constraints *Chun-Yi Su* +, *T-AC Mar 94* 609-614

structural wave control, reduced-order model *Quan Wang* + *T-AC Aug 94* 1711-1713

Mechanical variables control

superarticulated mech systs control *Seto D* + *I-AC Dec 94* 2442-2453

vibr systs control *Karl W C* + *T-AC Jan 94* 222-226

Mechanical variables control; cf. Force control, Motion control, Position control, Velocity control**Memoryless systems**

dyn lin time-delay systs, robust controller design *Mahmoud M S* + *T-AC May 94* 995-999

state delayed systs, memoryless H^∞ controllers *Joon Hwa Lee* + *T-AC Jan 94* 159-162

Wiener systs nonparametric ident *Greblicki W* *T-AC Oct 94* 2077-2086

Minimal realizations

full-order/low-order observers, LTI plant, unified singular syst based theory *Cobb J D* *T-AC Dec 94* 2497-2502

min stochastic realizations parameterization *Ferrante A* *T-AC Oct 94* 2122-2126

Minimax control

certainty equivalence control, dyn games *James M R* *T-AC Nov 94* 2321-2324

Minimax methods

optimal filtering, stochastic discrete-time systs, unknown inputs *Borisov A V* + *I-AC Dec 94* 2461-2464

Minimization methods

adaptive predictive controller robustness anal *Clarke D W* + *T-AC May 94* 1052-1056

least absol values estim computational aspects *Fedorov E D* *T-AC Mar 94* 626-630

lin quadratic control convex prog num method *Peres P L D* + *T-AC Jan 94* 198-202

mixed H_2/H_∞ perform objectives optimal control *Doyle J* + *T-AC Aug 94* 1575-1587

sampled-data robust control, exponential hold optimal algm *Yung-Chun Wu* + *I-AC Jan 94* 112-117

sampled-data systs wordlength constraint, stabil/perform *Fialho I J* + *I-AC Dec 94* 2476-2481

stochastic lin discrete systs state estim algm *Ahmed M S* *T-AC Aug 94* 1652-1656

uncertain syst robust stabil anal nonquadratic Lyapunov fns *Zelenitsky A L* *I-AC Jan 94* 135-138

uncertain systs stabilizing control design quasiconvex optim *Keqin Gu* *T-AC Jan 94* 127-131

Minimization methods; cf. Optimization methods**Minimum-energy control**

sampled data control systs H_∞ type problem *Hayakawa Y* + *I-AC Nov 94* 2278-2284

Mobile robotics, kinematics

kinematics product-of-exponentials formula computational aspects *Park I C* *T-AC Mar 94* 643-647

Mobile robots, motion planning

autonomous vision-based mobile robot *Baumgartner E T* + *T-AC Mar 94* 493-502

car reachable posns computation *Soueres P* + *T-AC Aug 94* 1626-1630

cat, falling, near-optimal nonholonomic motion planning, appl *Fernandes C* + *T-AC Mar 94* 450-463

trajectory stabilization for systs, nonholonomic constraints *Walsh G* + *T-AC Jan 94* 216-222

Modeling

2D syst model of Roesser, decomp *Afacan T* + *T-AC Nov 94* 2261-2262

A time-domain approach, model validation *Poolla K*, + *T-AC May 94* 951-959

continuous-time multivariable syst reduction *Krajewski W* + *T-AC Oct 94* 2126-2129

finite-dimens model validation, output error, test horizon *Hoai Nghia Duong* + *T-AC Jan 94* 102-106

Model reduction; cf. Reduced-order systems**Model reference adaptive control**

design for plants, unknown dead-zones *Gang Tao* + *T-AC Jan 94* 59-68

multiple models/switching approach *Narendra K S* + *T-AC Sep 94* 1861-1866

multivariable direct MRAC, identifiable parameterizations and param convergence *de Mathelin M* + *T-AC Aug 94* 1612-1617

multivariable lin systs, hysteresis switching MRAC *Weller, S R*, + *T-AC Jul 94* 1360-1375

parab systs direct adaptive control *Keum Shik Hong*, + *T-AC Oct 94* 2018-2033

perform anal /improvement, new tracking error criteria *Dattam A* +, *T-AC Dec 94* 2370-2387

robust MRAC SISO systs *Zhihua Qu* + *T-AC Nov 94* 2219-2234

robust stabilization/perform, MRAC modeling error compensation *Sun J* + *T-AC Mar 94* 630-635

sing-free multivariable MRAC *Moctezuma R G* + *T-AC Sep 94* 1856-1860

stochastic adaptive prediction/MRAC *Wei Ren* + *T-AC Oct 94* 2047-2060

suboptimal perfect model matching, with noise, MRACS *Mutoh Y* + *T-AC Feb 94* 422-425

transient perform improvement by feedback *Datta A* + *I-AC Sep 94* 1977-1980

variable struct MRAC, I/O based, anal /design *Liu Hsu* + *T-AC Jan 94* 4-21

Motion control

constrained robots, force/motion control *Grabbe M T* + *T-AC Jan 94* 179

robot variable struct control schemes *Bin Yao* + *I-AC Feb 94* 371-376

Motion control; cf. Manipulators, Mobile robots**Motion measurement; cf. Tracking****Motion planning**

cat, falling, near-optimal nonholonomic motion planning, appl *Fernandes C* + *T-AC Mar 94* 450-463

Motors; cf. DC motors**Moving-average processes; cf. Autoregressive moving-average processes****MRAC; cf. Model reference adaptive control****Multidimensional systems**

evol eqns, parab partial DE, exponential stabil *Keum-Shik Hong* + *T-AC Jul 94* 1432-1436

finite-dimens controllers designed for infinite-dimens systs state space *Morris K A* *T-AC Oct 94* 2100-2104

model validation, output error, test horizon *Hoai Nghia Duong* + *I-AC Jan 94* 102-106

neutral systs, stabil and stabilizability *Logemann H* + *T-AC Jan 94* 138-143

nonnegative feedback control oscills stabilisation *Zaslavsky B* *T-AC Jun 94* 1273-1276

parab systs direct adaptive control *Keum Shik Hong* + *T-AC Oct 94* 2018-2033

robust adaptive pole placement control *Weyer E* + *T-AC Aug 94* 1665-1671

spectral syst finite-dimens approx *Erickson M A* + *T-AC Sep 94* 1904-1909

stabilization, adaptive, nonlin time-varying controller *Miller D E* *T-AC Jul 94* 1347-1359

stochastic dyn systs exact lin *Socha L* *T-AC Sep 94* 1980-1984

worst-case ident anal, BIBO robustness *Partington J R* + *T-AC Oct 94* 2171-2176

Multinput-multiooutput systems; cf. Multivariable systems**Multilinear systems; cf. Bilinear systems****Multiplication; cf. Matrix multiplication****Multisensor systems**

decentralized detect, optimal thresholds *Irving W W* + *T-AC Apr 94* 835-838

detect networks with multiple event struct, optim *Pete A* + *T-AC Aug 94* 1702-1707

multiresolutional distributed filtering *Lang Hong* *T-AC Apr 94* 853-856

multiscale recursive estim, data fusion/regularization *Chou K C* + *T-AC Mar 94* 464-478

Multivariable systems

book review, Linear Multivariable Control Algebraic Analysis and Synthesis Methods (Vardoulakis, A I G, 1991) *Lewis F L* *T-AC Jul 94* 1536

comments on "Stability margin evaluation for uncertain linear systems" (by C Gong and S Thompson, Jun 94 548-550) *Su J-H* *T-AC Dec 94* 2523-2524

continuous-time adaptive decoupling control design *Ortega R* + *T-AC Aug 94* 1639-1643

continuous-time multivariable syst reduction *Krajewski W* + *T-AC Oct 94* 2126-2129

cyclicly switched param-adaptive control systs, MIMO design models/internal regulators *Morse, A S* + *T-AC Sep 94* 1809-1818

decentralized control, closed loop props *Campo P J* + *T-AC May 94* 932-943

diagonal decoupling, dyn. output feedback/const. precompensator. *Eldem, V.*, T-AC Mar 94 503-511
 direct MRAC algn. *de Matheltn, M.*, +, T-AC Aug 94 1612-1617
 eigenstructure assignment by decentralized output feedback. *Guang-Ren Duan*, T-AC May 94 1009-1014
 floating platform, modeling and control. *Damen, A.A.H.*, +, T-AC May 94 1075-1078
 H_∞ model reduction computational scheme. *Kavranoglu, D.*, T-AC Jul 94 1447-1451
 hysteresis switching MRAC of lin. multivariable systs. *Weller, S.R.*, +, T-AC Jul 94 1360-1375
 ident., state-space freq. domain approach. *Bayard, D.S.*, T-AC Sep 94 1880-1885
 large scale interconnected systs. with delays, decentralized stabilization. *Zhongzhi Hu*, T-AC Jan 94 180-182
 lin. multivariable systs., fund. notion of equivalence. *Pugh, A.C.*, +, T-AC May 94 1141-1145
 MIMO lin. systs. with sens./actuator failures, stabil. *Gundes, A.N.*, T-AC Jun 94 1224-1230
 MIMO nonlin. syst. I/O pseudolinearization. *Lawrence, D.A.*, +, T-AC Nov 94 2207-2218
 MIMO stochastic discrete syst. state estim. *Liu Danyang*, +, T-AC Oct 94 2087-2091
 multichannel gain margin improvement, sampled-data hold fns. *Chang Yang*, +, T-AC Mar 94 657-661
 multivariable nonminimum phase discrete-time systs. LTR procedure. *Leon de la Barra S., B.A.*, T-AC Mar 94 574-577
 multivariable systs., quantitat. robustness measures. *Hong-Giou Chen*, +, T-AC Apr 94 807-810
 nonlin. act. vibr. damping. *Kanestrom, R.K.*, +, T-AC Sep 94 1925-1928
 nonlin. systs., asymptotic tracking, necessary conditions. *Grizzle, J.W.*, +, T-AC Sep 94 1782-1794
 output feedback problem in lin. systs., bilinear formulation. *Syrmos, V.L.*, +, T-AC Feb 94 410-414
 robust anal., uncertainty value sets. *Eszter, E.G.*, +, T-AC Nov 94 2315-2318
 robust MIMO terminal sliding-mode control, rigid robotic manipulators. *Zhihong, M.*, +, T-AC Dec 94 2464-2469
 sing.-free multivariable MRAC. *Moctezuma, R.G.*, +, T-AC Sep 94 1856-1860
 std. multivariable H₂-optimal control problem, polynomial soln. *Hunt, K.J.*, +, T-AC Jul 94 1502-1507
 switching surface design for multivariable VSS. *Ju-Jang Lee*, +, T-AC Feb 94 414-419
 uncertain lin. systs., multivariable, stabil. margin eval. *Gong, C.*, +, T-AC Mar 94 548-550
 uncertain nonlin. systs., struct. invariance. *Castro-Linares, R.*, +, T-AC Oct 94 2154-2158

N

Navigation; cf. Land navigation

Networks; cf. Petri nets

Neural network applications

binary hypothesis testing, struct. adaptive nets. *Papadakis, I.N.M.*, +, T-AC Sep 94 1967-1971
 quantized constrained systs., feedback control, appls. to neuromorphic controllers design. *Sznaier, M.*, +, T-AC Jul 94 1497-1502

Neurocontrollers

book review; Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches (White, D.A., and Sofge, D.A., Eds.; 1992). *Samad, T.*, T-AC Jul 94 1534-1535
 continuous-time nonlin. systs. adaptive control by neural nets. *Fu-Chuang Chen*, +, T-AC Jun 94 1306-1310
 quantized constrained systs., feedback control, appls. to neuromorphic controllers design. *Sznaier, M.*, +, T-AC Jul 94 1497-1502

Newton's method

dyn. shock-error models, online estim. *Krishnamurthy, V.*, T-AC May 94 1129-1135
 state-constrained optimal control, generalised dual quasi-Newton algn. *Shimizu, K.*, +, T-AC May 94 982-986

Noise

continuous time systs., transfer fns., model errors. *Schoukens, J.*, +, T-AC Aug 94 1733-1737
 failure detect./isolation/accommodation syst. *Chia-Chi Tsui*, T-AC Nov 94 2318-2321
 model matching, suboptimal perfect, with noise, MRACS. *Mutoh, Y.*, +, T-AC Feb 94 422-425
 worst-case syst. ident., time complexity. *Poolla, K.*, +, T-AC May 94 944-950

Noise; cf. Filter noise; Gaussian noise; White noise

Nonlinear difference equations; cf. Riccati equations, discrete-time

Nonlinear differential equations

adaptive control, nonlin. systs., triangular struct. *Seto, D.*, +, T-AC Jul 94 1411-1428
 almost sure sample stabil. of nonlin. stochastic dyn. systs. *Zhi Yu Zhang*, +, T-AC Mar 94 560-565
 sing. perturbed control systs. and nonlin. differential-alg. eqns. *Krishnan, H.*, +, T-AC May 94 1079-1084

Nonlinear differential equations; cf. Riccati equations, differential

Nonlinear equations; cf. Newton's method; Riccati equations, algebraic

Nonlinear filtering

nonlin. syst. state estim., conditionally min. algn. *Pankov, A.R.*, +, T-AC Aug 94 1617-1620

Nonlinearities; cf. Hysteresis nonlinearities

Nonlinear systems

adaptive control systs., transient bounds. *Zhuquan Zang*, +, T-AC Jan 94 171-175
 almost sure sample stabil. of nonlin. stochastic dyn. systs. *Zhi Yu Zhang*, +, T-AC Mar 94 560-565
 asymptotic model matching. *Di Benedetto, M.D.*, +, T-AC Aug 94 1539-1550
 asymptotic tracking, necessary conditions. *Grizzle, J.W.*, +, T-AC Sep 94 1782-1794
 bilinear systs., state space anal., orthogonal series approach. *Paraskevopoulos, P.N.*, +, T-AC Apr 94 793-797
 book review; Nonlinear System Analysis, 2nd edn. (Vidyasagar, M., 1993). *Abed, E.H.*, T-AC Jul 94 1535-1536
 book review; Nonlinear Systems (Khalil, H.; 1992). *Grizzle, J.W.*, T-AC Jan 94 251-252
 brushless DC motors, feedback-lin. control charactn. *In-Joong Ha*, +, T-AC Mar 94 673-677
 cascaded nonlin. systs., global robust stabilization. *Imura, J.-I.*, +, T-AC May 94 1084-1089
 certainty equivalence control, dyn. games. *James, M.R.*, T-AC Nov 94 2321-2324
 constrained nonlin. syst. receding horizon control. *Shin-Yeu Lin*, T-AC Sep 94 1893-1899
 control algn., automatic differentiation appl. *Campbell, S.L.*, +, T-AC May 94 1047-1052
 delay-independent exponential stabil. criteria for time-varying discrete delay systs. *Wu, J.W.*, +, T-AC Apr 94 811-814
 detect. networks with multiple event struct., optim. *Pete, A.*, +, T-AC Aug 94 1702-1707
 deterministic nonlin. syst. ident. *Shir-Kuan Lin*, T-AC Sep 94 1886-1893
 differential-delay systs., absol. stabil. *Gil, M.I.*, T-AC Dec 94 2481-2484
 discrete-time adaptive nonlin. syst., least sq. estimator. *Kanellakopoulos, I.*, T-AC Nov 94 2362-2365
 discrete time Hammerstein syst. ident. *Lang Zi-Qiang*, T-AC Mar 94 569-573
 discrete-time nonlin. control systs., design via smooth feedback. *Wei Lin*, +, T-AC Nov 94 2340-2346
 discrete-time nonlin. syst. control, stabil. property. *Jie Huang*, +, T-AC Nov 94 2307-2311
 discrete time nonlin. systs., stabil., time delayed feedback. *Xueshan Yang*, +, T-AC Mar 94 585-589
 discrete-time systs., input-to-state stabil. condition, global stabilization. *Kazakos, D.*, +, T-AC Oct 94 2111-2113
 first-order nonlin. syst. adaptive control. *Brogliato, B.*, +, T-AC Aug 94 1764-1768
 fuzzy control systs., supervisory controller. *Li-Xiu Wang*, T-AC Sep 94 1845-1847
 H_∞ control, output-feedback based. *Lu, W.-M.*, +, T-AC Dec 94 2517-2524
 input-output decoupling, nonlin. interactor appl. *Di Benedetto, M.D.*, +, T-AC Jun 94 1246-1250
 input-output lin., state equivalence and decoupling. *In-Joong Ha*, +, T-AC Nov 94 2269-2274
 internally stable nonlin. systs. with disturbances, parameterization. *Hammer, J.*, T-AC Feb 94 300-314
 interval plants, closed-loop hyperstability. *Foo, Y.K.*, +, T-AC Jan 94 151-154
 lin. systs., nonlin. design of adaptive controllers. *Krstic, M.*, +, T-AC Apr 94 738-752
 MIMO nonlin. syst. I/O pseudolinearization. *Lawrence, D.A.*, +, T-AC Nov 94 2207-2218
 MRAC, robust, SISO systs. *Zhihua Qu*, +, T-AC Nov 94 2219-2234
 multi-DOF nonlin. damping model, spectral dens. *Wei Jian Zhang*, T-AC Feb 94 406-410
 nonlin. servomechanism robust control. *Jie Huang*, +, T-AC Jun 94 1510-1513
 nonnegative feedback control, oscills. stabilisation. *Zaslavsky, B.*, T-AC Jun 94 1273-1276

- num design *Kreisselmeier G* + *T-AC Jan 94* 33-46
 output feedback stabilization of nonlin systs *Tsimas J* + *T-AC Apr 94* 806
 parametrically uncertain nonlin feedback systs, absol stabil *Marquez H J* + *T-AC Mar 94* 664-668
 parametrically uncertain nonlin systs, robust stabilisation *Schoenwald D A* + *T-AC Aug 94* 1751-1755
 partially obs discrete-time nonlin systs, risk-sensitive control and dyn games *James M R* + *T-AC Apr 94* 780-792
 pos real systs nonlin controllers *Bernstein D S* + *T-AC Jul 94* 1513-1517
 P-type learning control *Saab S S* *T-AC Nov 94* 2298-2302
 recursive ident, nonlin Wiener model *Wigren T* *T-AC Nov 94* 2191-2206
 robots, trajectory stabilization for systs, nonholonomic constraints *Walsh G* + *T-AC Jan 94* 216-222
 robust control design, cascade struct approach *Bonivento C* + *T-AC Apr 94* 846-849
 stabilization, adaptive, nonlin time-varying controller *Miller D E* *T-AC Jul 94* 1347-1359
 stabilization of discrete-time nonlin systs global *Byrnes C I* + *T-AC Jan 94* 83-98
 state estim, conditionally min nonlin filtering *Pankov A R* + *T-AC Aug 94* 1617-1620
 stochastic dyn systs, exact lin *Socha L* *T-AC Sep 94* 1980-1984
 switched-mode power converters multirate modeling and control design *Khayatian A* + *T-AC Sep 94* 1848-1852
 time lag nonlin systs, vibr control *Lehman B* + *T-AC May 94* 898-912
 time-varying functional diff eqns, delay independent stabil conditions/decay estim *Lehman B* + *T-AC Aug 94* 1673-1676
 uncertain nonlin interconnected systs time-varying state delay, stabilizing control *Mahmoud M S* *T-AC Dec 94* 2484-2488
 uncertain nonlin systs struct invariance *Castro Linares R* + *T-AC Oct 94* 2154-2158
 uncertain nonlin systs tracking control *Song Y D* + *T-AC Sep 94* 1866-1871
 uncertain SISO min phase lin syst nonlin universal servomechanism *Ryan E P* *T-AC Apr 94* 753-761
 vibr control by AP-forcing *Balestrino A* + *T-AC Jun 94* 1255-1258
 Wiener systs nonparametric ident *Greblicki W* *T-AC Oct 94* 2077-2086
- Nonlinear systems, cf Bilinear systems**
Nonparametric estimation
 Wiener systs nonparametric ident *Greblicki W* *T-AC Oct 94* 2077-2086
- Numerical integration**
 Routh table prop quadratic integral eval appl *Beghi A* + *T-AC Dec 94* 2494-2496
- Numerical methods**
 dyn syst simul fast parallel recursive aggregation *Tsai W A* + *T-AC Mar 94* 534-540
 greatest common divisor of polynomials comput matrix pencil based num method *Karcanias N* + *T-AC May 94* 977-981
 lin quadratic control convex prog num method *Peres P L D* + *T-AC Jan 94* 198-202
- Numerical methods, cf Approximation methods** Convergence of numerical methods, Newton's method Optimization methods, Relaxation methods
- Nyquist stability**
 closed loop syst freq response plots comment *Gessing R* *T-AC Aug 94* 1770-1771
 interval lin control syst robust parametric design *Keel L H* + *T-AC Jul 94* 1524-1530
 interval plant family Nyquist envelope *Hollot C V* + *T-AC Feb 94* 391-396
 parametrically uncertain systs, anal and synthesis *Kaminsky R D* + *T-AC Apr 94* 874-876
- O**
- Observability**
 continuous descriptor systs adaptive observer *Uetake Y* *T-AC Oct 94* 2095-2100
 LTI, observability delay *Bose S* + *T-AC Apr 94* 803-806
 LTI, output back control by multirate PWM *Khayatian A* + *T-AC Jun 94* 1292-1297
 LTI, stochastic control syst tuning *van Schuppen J H* *T-AC Nov 94* 178-190
 LTI, discrete-time, Kalman filters, Riccati eqns *Chou K C* + *T-AC Mar 94* 479-492
 LTI, observer, state feedback, asymptotic stabil *Hunt L R* + *T-AC Oct 94* 2113-2118
 LTI, realness conditions charactn assuming controllability *Weiss H* + *T-AC Mar 94* 540-544
- SISO LTI discrete-time systs, param-adaptive controller *Kreisselmeier G* *T-AC Sep 94* 1819-1826
 switched-mode power converters, multirate modeling and control design *Khayatian A* + *T-AC Sep 94* 1848-1852
 time-invariant lin systs controllability/observability *Kaining Wang* + *T-AC Jul 94* 1443-1447
- Observers**
 asynchronous sampled data syst control design *Voulgaris P* *T-AC Jul 94* 1451-1455
 DESS, observability, delay *Bose S* + *T-AC Apr 94* 803-806
 discrete-time syst exponential stabil observer design *Aitken V C* + *T-AC Sep 94* 1959-1962
 disturbance decoupled observer design *Hou M* + *T-AC Jun 94* 1338-1341
 full-order/low-order observers, LTI plant, unified singular syst based theory *Cobb J D* *T-AC Dec 94* 2497-2502
 H_∞ compensator design min order observers *Hsu C S* + *T-AC Aug 94* 1679-1681
 H_∞ -optim, reduced order observer based controller *Stoorvogel A A* + *T-AC Feb 94* 355-360
 lin observers, parameterization and design *Ding X* + *T-AC Aug 94* 1648-1652
 multibody flexible systs elastic mode estim *Karray F* + *T-AC May 94* 1016-1020
 observer design, eigenvalue assignment, residual generation *Magni J-F* + *T-AC Feb 94* 441-447
 observers for nonlin systs in steady state *Hunt L R* + *T-AC Oct 94* 2113-2118
 partially obs discrete-time nonlin systs, risk-sensitive control and dyn games *James M R* + *T-AC Apr 94* 780-792
 robots, robust control, lin estim state feedback *Berghuis H* + *T-AC Oct 94* 2159-2162
 slowly varying and LTI systs, ident and uncertainty principles *Zames G* + *T-AC Sep 94* 1827-1838
 state observer/feedback compensators, comment *Bender D J* *T-AC Feb 94* 447-448
 switched-mode power converters, multirate modeling and control design *Khayatian A* + *T-AC Sep 94* 1848-1852
 unknown input lin systs full-order observer design *Darouach M* + *T-AC Mar 94* 606-609
 unknown input observers design *Darouach M* *T-AC Mar 94* 698-699
 unstable plants sampled-data observers generalized holds *Haddad W M* + *T-AC Jan 94* 229-234
- Observers, cf Adaptive observers**
- Optimal control**
 bilinear systs, adaptive stabilization least sq *Xi Sun* + *T-AC Jan 94* 207-211
 book review 'Applied Optimal Control and Estimation' (Lewis F L 1992) *Kamen E W* *T-AC Aug 94* 1773-1774
 certainty equivalence control dyn games *James M R* *T-AC Nov 94* 2321-2324
 closed Jackson queueing network decentralized control *Dye Jyun Ma* + *T-AC Jul 94* 1460-1463
 descriptor systs lin feedback closed-loop Stackelberg strategy *Hua Xu* + *T-AC May 94* 1097-1102
 detect networks with multiple event struct, optim *Pete A* + *T-AC Aug 94* 1702-1707
 discrete-time Riccati eqn H_∞ control appl *Stoorvogel A A* + *T-AC Mar 94* 686-691
 discrete-time systs mixed H_2/H_∞ control exact convex optim based soln *Sznajder M* *T-AC Dec 94* 2511-2517
 dissipative H_2/H_∞ controller synthesis *Haddad W M* + *T-AC Apr 94* 827-831
 failure-prone mfg systs, structural props of optimal prod controllers *Jian-Qiang Hu* + *T-AC Mar 94* 640-643
 finite-time optimal control of bilinear systs successive approx procedure *Aganovic Z* + *T-AC Sep 94* 1932-1935
 floating platform, modeling and control *Damen A A H* + *T-AC May 94* 1075-1078
 four-block problem, recursive Schur-based soln *Constantinescu T* + *T-AC Jul 94* 1476-1481
 H_∞ control syst ident, least sq methods, comment *Livstone M M* + *T-AC Jul 94* 1531
 H_∞ model reduction computational scheme *Kavranoglu D* *T-AC Jul 94* 1447-1451
 H_2 -optimal zeros placement *Kabamba P T* + *T-AC Jun 94* 1298-1301
 H_∞ control of discrete-time uncertain systs *Geromel J C* + *T-AC May 94* 1072-1075
 H_∞ control of systs under norm bounded uncertainties *Gu K* *T-AC Jun 94* 1320-1322
 H_∞ -optimal discrete-time fixed-point and fixed-lag smoothing, game theory approach *Theodor Y* + *T-AC Sep 94* 1944-1948

- H^∞ -optimal control for sing. perturbed systs., imperfect state meas. *Zigang Pan, +*, T-AC Feb 94 280-299
- infinite horizon LQ control, soln. approx. *Schochetman, I.E., +*, T-AC Mar 94 596-601
- L^2 optimal model order reduction problem, input normal form homotopy. *Ge, Y., +*, T-AC Jun 94 1302-1305
- mixed H_2/H_∞ control, Nash game approach. *Limebeer, D.J.N., +*, T-AC Jan 94 69-82
- M/M/1 queue with bal. budget, optimal flow control. *Chakravorti, B., +*, T-AC Sep 94 1918-1921
- multirate sampled-data systs., H_2 -optimal design. *Qui, L., +*, T-AC Dec 94 2506-2511
- optimal dyn. output feedback for nonzero set point regulation, discrete-time case. *Haddad, W.M., +*, T-AC Sep 94 1921-1925
- optimal filtering, stochastic discrete-time systs., unknown inputs. *Borisov, A.V., +*, T-AC Dec 94 2461-2464
- periodic discrete-time Riccati eqn., num. soln. *Hench, J.J., +*, T-AC Jun 94 1197-1210
- prod. control, multiple time scales approach. *Jiang, J., +*, T-AC Nov 94 2292-2297
- sampled-data robust control, exponential hold, optimal algm. *Yung-Chun Wu, +*, T-AC Jan 94 112-117
- SISO-distributed plants, optimal mixed sensitivity. *Flamm, D.S., +*, T-AC Jun 94 1150-1165
- state-constrained optimal control, generalised dual quasi-Newton algm. *Shimizu, K., +*, T-AC May 94 982-986
- std. multivariable H_2 -optimal control problem, polynomial soln. *Hunt, K.J., +*, T-AC Jul 94 1502-1507
- stochastic dyn. job shops/prod. planning. *Sethi, S., +*, T-AC Oct 94 2061-2076
- transfer matrices preconditioning. *Rotstein, B., +*, T-AC Nov 94 2287-2292
- vector discrete-event syst. controller synthesis. *Yong Li, +*, T-AC Mar 94 512-531
- Optimal control; cf. Cost-optimal control; H^∞ optimization; Linear-quadratic control; Minimax control; Minimum-energy control; Quantized control; Stochastic optimal control; Suboptimal control; Time-optimal control**
- Optimization methods**
- constrained optim. in Hilbert space. *Shimizu, K., +*, T-AC May 94 982-986
- data network distributed asynchronous routing, convergence. *Zhi-Quan Luo, +*, T-AC May 94 1123-1129
- data network path formulated optimal routing, time complexity. *Antonio, J.K., +*, T-AC Feb 94 385-391
- detect. networks with multiple event struct., optim. *Pete, A., +*, T-AC Aug 94 1702-1707
- extended horizon liftings for nonminimum-phase systs. *Bayard, D.S., +*, T-AC Jun 94 1333-1338
- failure prone mfg. syst. hedging point policies. *Jian-Qiang Hu, +*, T-AC Sep 94 1875-1880
- global optim., adaptive partitioned random search. *Bo Tang, Z., +*, T-AC Nov 94 2235-2244
- heuristics, rules of thumb, and the 80/20 proposition. *Yu-Chi Ho, +*, T-AC May 94 1025-1027
- H_∞ control syst. ident., least sq. methods, comment. *Livstone, M.M., +*, T-AC Jul 94 1531
- ident., family of norms. *Massoumnia, M.-A., +*, T-AC May 94 1027-1031
- Kalman filtering for uncertain discrete-time systs. *Lihua Xie, +*, T-AC Jun 94 1310-1314
- lin. syst. approx., time-scaling factor, Laguerre models. *Wang, L., +*, T-AC Jul 94 1463-1467
- L^2 optimal model order reduction problem, input normal form homotopy. *Ge, Y., +*, T-AC Jun 94 1302-1305
- MIMO stochastic discrete syst. state estim. *Liu Danyang, +*, T-AC Oct 94 2087-2091
- min. lin. plants parameterization. *Davis, L.D., +*, T-AC Apr 94 849-852
- multiresolutional distributed filtering. *Lang Hong, +*, T-AC Apr 94 853-856
- path formulated optimal routing algm., time complexity. *Antonio, J.K., +*, T-AC Sep 94 1839-1844
- perturbed lin. systs., upper covariance bounds. *Bolzern, P., +*, T-AC Mar 94 623-626
- rational transfer fn., optimal L_∞ approx. *Kavranoglu, D., +*, T-AC Sep 94 1899-1904
- regenerative syst. optim., infinitesimal perturb. anal. *Chong, E.K.P., +*, T-AC Jul 94 1400-1410
- sens. fusion, decentralized detect. optimal thresholds. *Irving, W.W., +*, T-AC Apr 94 835-838
- single-machine scheduling, resource optimal control. *Cheng, T.C.E., +*, T-AC Jun 94 1243-1246
- SISO discrete-time systs. preview tracking. *Halpern, M.E., +*, T-AC Mar 94 589-592
- SISO systs., robust stabilization. *Olbrot, A.W., +*, T-AC Mar 94 652-657
- time-invariant lin. systs., robust controller synthesis. *Rantzer, A., +*, T-AC Sep 94 1802-1808
- uncertain syst. pole assignment, polynomial approach. *Figueroa, J.L., +*, T-AC Apr 94 831-835
- Optimization methods; cf. Approximation methods; Gradient methods; Least-mean-square methods; Least-squares methods; Minimization methods; Minimization methods**
- Output feedback**
- closed-loop vibr. control, Youla parameterization. *Kabamba, P.T., +*, T-AC Jul 94 1455-1459
- comments on "Stabilization via static output feedback". *Pimpalkhare, A.A., +*, T-AC May 94 1148
- continuous-time nonlin. systs. adaptive control by neural nets. *Fu-Chuang Chen, +*, T-AC Jun 94 1306-1310
- descriptor systs. control by output feedback. *Lovass-Nagy, V., +*, T-AC Jul 94 1507-1509
- descriptor systs. regularization by output feedback. *Bunse-Gerstner, A., +*, T-AC Aug 94 1742-1748
- diagonal decoupling, dyn. output feedback/const. precompensator. *Eldem, V., +*, T-AC Mar 94 503-511
- discrete time nonlin. systs., stabil., time delayed feedback. *Xueshan Yang, +*, T-AC Mar 94 585-589
- eigenstructure assignment by decentralized output feedback. *Guang-Ren Duan, +*, T-AC May 94 1009-1014
- feedback control by multirate PWM. *Khayatian, A., +*, T-AC Jun 94 1292-1297
- H_∞ -control, state availability rel., pole/zero cancellations. *Miyamoto, S., +*, T-AC Feb 94 379-381
- MIMO lin. systs. with sens./actuator failures, stabil. *Gundes, A.N., +*, T-AC Jun 94 1224-1230
- mobile manipulator locomotion/manipulation coord. *Yamamoto, Y., +*, T-AC Jun 94 1326-1332
- nonlin. syst. H_∞ control, output-feedback based. *Lu, W.-M., +*, T-AC Dec 94 2517-2524
- output feedback problem in lin. systs., bilinear formulation. *Syrmos, V.L., +*, T-AC Feb 94 410-414
- output feedback stabilization of nonlin. systs. *Tsinias, J., +*, T-AC Apr 94 806
- pole placement by output feedback. *Lee, T.H., +*, T-AC Mar 94 565-568
- pos. realness conditions, charactn. assuming controllability. *Weiss, H., +*, T-AC Mar 94 540-544
- stabilization of discrete-time nonlin. systs., feedback equivalence. *Byrnes, C.I., +*, T-AC Jan 94 83-98
- time-invariant lin. syst. adaptive control. *Karason, S.P., +*, T-AC Nov 94 2325-2330
- uncertain nonlin. interconnected systs., time-varying state delay, stabilizing control. *Mahmoud, M.S., +*, T-AC Dec 94 2484-2488
- uncertain systs. stabilizing control design, quasiconvex optim. *Keqin Gu, +*, T-AC Jan 94 127-131
- Packet radio**
- optimal multicopy Aloha. *Wong, E.W.M., +*, T-AC Jun 94 1233-1236
- Packet switching**
- flow control, rate-based, monotonicity/concavity. *Budka, K.C., +*, T-AC Mar 94 544-548
- optimal multicopy Aloha. *Wong, E.W.M., +*, T-AC Jun 94 1233-1236
- Parallel processing**
- constrained nonlin. syst. receding horizon control. *Shin-Yeu Lin, +*, T-AC Sep 94 1893-1899
- Parameter estimation**
- adaptive control by multiple models/switching. *Narendra, K.S., +*, T-AC Sep 94 1861-1866
- adaptive control systs., chaotic behavior. *Gonzalez, G.A., +*, T-AC Oct 94 2145-2148
- adaptive pole placement without excitation probing sigs. *Lozano, R., +*, T-AC Jan 94 47-58
- autonomous vision-based mobile robot. *Baumgartner, E.T., +*, T-AC Mar 94 493-502
- bounded-error tracking of time-varying params. *Piet-Lahanier, H., +*, T-AC Aug 94 1661-1664
- discrete-time adaptive nonlin. syst., least sq. estimator. *Kanellakopoulos, I., +*, T-AC Nov 94 2362-2365
- extended Chandrasekhar recursions. *Sayed, A.H., +*, T-AC Mar 94 619-623
- first-order nonlin. syst. adaptive control. *Brogliato, B., +*, T-AC Aug 94 1764-1768
- guaranteed param. estim. problem with uncertain stats., Kalman filter accuracy. *Matasov, A.I., +*, T-AC Mar 94 635-639
- least absol. values estim., computational aspects. *Flodorov, E.D., +*, T-AC Mar 94 626-630

- least sq. estim. in white noise, convergence. *Nassiri-Toussi, K.*, +, *T-AC Feb 94* 364-368
- lin. syst. approx., time-scaling factor, Laguerre models *Wang, L.*, +, *T-AC Jul 94* 1463-1467
- manipulators with parametric uncertainty, robust control *Keun-Mo Koo*, +, *T-AC Jun 94* 1230-1233
- multibody flexible systs., elastic mode estim. *Karray, F.*, +, *T-AC May 94* 1016-1020
- multivariable direct MRAC, identifiable parameterizations and param. convergence. *de Mathelin, M.*, +, *T-AC Aug 94* 1612-1617
- nonminimum phase first-order continuous-time systs., adaptive stabilization *Lozano, R.*, +, *T-AC Aug 94* 1748-1751
- random param tracking, robust algm *Juditsky, A.*, +, *T-AC Jun 94* 1211-1221
- stochastic lin. syst. param hypothesis testing, stat sampling *Duncan, T.E.*, +, *T-AC Jan 94* 118-122
- time delay estim. in continuous LTI systs *Tuch, J.*, +, *T-AC Apr 94* 823-827
- Parameter identification**
- continuous time systs., transfer fns., model errors *Schoukens, J.*, +, *T-AC Aug 94* 1733-1737
- deterministic nonlin. syst. ident. *Shir-Kuan Lin*, *T-AC Sep 94* 1886-1893
- time-varying params., bounded error ident. *Bittanti, S.*, +, *T-AC May 94* 1106-1110
- transfer fn. param. ident. in freq. domain *Pintelon, R.*, +, *T-AC Nov 94* 2245-2260
- Parameter identification; cf. Power system identification**
- Parameter-space methods**
- cyclicly switched param.-adaptive control systs., MIMO design models/ internal regulators *Morse, A.S.*, +, *T-AC Sep 94* 1809-1818
- H_∞ control of discrete-time uncertain systs. *Geromel, J.C.*, +, *T-AC May 94* 1072-1075
- param.-adaptive control, cyclic switching strategy *Pati, F.M.*, +, *T-AC Jun 94* 1172-1183
- Parameter uncertainty; cf. Uncertain systems**
- Partial differential equations**
- evol. eqns., parab. partial DE, exponential stabil. *Keum-Shik Hong*, +, *T-AC Jul 94* 1432-1436
- parab. systs. direct adaptive control *Keum Shik Hong*, +, *T-AC Oct 94* 2018-2033
- spectral syst. finite-dimens. approx. *Erickson, M.A.*, +, *T-AC Sep 94* 1904-1909
- Passivity**
- rigid spacecraft adaptive attitude control *Egeland, O.*, +, *T-AC Apr 94* 842-846
- robot adaptive control based on passivity *Yu Fang*, +, *T-AC Sep 94* 1871-1875
- second-order dyn. systs., dissipative controller designs *Morris, K.A.*, +, *T-AC May 94* 1056-1063
- Periodic control**
- multirate sampled-data systs. struct. props. *Longhi, S.*, *T-AC Mar 94* 692-696
- periodically time-varying lin. discrete-time plants, decentralized control *Khargonekar, P.P.*, +, *T-AC Apr 94* 877-882
- Perturbation methods**
- discrete-time systs., perturb., comment *Eslami, M.*, *T-AC Aug 94* 1768-1769
- regenerative syst. optim., infinitesimal perturb. anal. *Chong, E.K.P.*, +, *T-AC Jul 94* 1400-1410
- Petri nets**
- event graphs, stochastic, cycle time *Proth, J.-M.*, +, *T-AC Jul 94* 1482-1486
- logical discrete event systs., Ilyapunov stabil. *Passino, K.M.*, +, *T-AC Feb 94* 269-279
- supervisory control, blocking and controllability of Petri nets. *Giua, A.*, +, *T-AC Apr 94* 818-823
- Planning; cf. Manipulators, motion planning; Manufacturing planning, Mobile robots, motion planning**
- Poisson processes**
- asynchronous systs. with Poisson transits, stabil. *Leland, R.P.*, *T-AC Jan 94* 182-185
- Pole assignment**
- adaptive pole placement without excitation probing sigs. *Lozano, R.*, +, *T-AC Jan 94* 47-58
- disturbance decoupling, constrained Sylvester eqns. *Syrmos, V.L.*, *T-AC Apr 94* 797-803
- D-stabil., robust, for lin. uncertain discrete delay systs. *Te-Jen Su*, +, *T-AC Feb 94* 425-428
- Gaussian stochastic control syst. tuning *van Schuppen, J.H.*, *T-AC Nov 94* 2178-2190
- n. discrete syst. pole assignment *Benzaouia, A.*, *T-AC Oct 94* 2091-2095
- pole placement by output feedback *Lee, T.H.*, +, *T-AC Mar 94* 565-568
- robust adaptive pole placement control. *Weyer, E.*, +, *T-AC Aug 94* 1665-1671
- uncertain continuous-time implicit systs., regional pole placement, robustness. *Chun-Hsiung Fang*, +, *T-AC Nov 94* 2303-2307
- uncertain syst. pole assignment, polynomial approach. *Figuerou, J.L.*, +, *T-AC Apr 94* 831-835
- variable struct. generator voltage regulator design, pole assignment tech. *Aggoune, M.E.*, +, *T-AC Oct 94* 2106-2110
- Poles and zeros**
- adaptive pole placement without excitation probing sigs. *Lozano, R.*, +, *T-AC Jan 94* 47-58
- comments on "System zeros determination from an unreduced matrix fraction description" (by K.S. Yeung and C.-M. Kwan, Nov 93 1695-1697) *Ferreira, P.M.G.*, *T-AC Nov 94* 2367
- corrections to "On undershoot in SISO systems" (Mar 94 578-581) *de la Barra S., B.A.L.*, *T-AC Aug 94* 1771
- D-stabil., robust, for lin. uncertain discrete delay systs. *Te-Jen Su*, +, *T-AC Feb 94* 425-428
- dyn. lin. systs., pos. real lemma generalization *Scherer, R.*, +, *T-AC Apr 94* 882-886
- extended horizon liftings for nonminimum-phase systs. *Bayard, D.S.*, *T-AC Jun 94* 1333-1338
- H_2 -optimal zeros placement *Kabamba, P.T.*, +, *T-AC Jun 94* 1298-1301
- H_∞ -control, state availability rel., pole/zero cancellations *Miyamoto, S.*, +, *T-AC Feb 94* 379-381
- iter. algm. for pole placement *Lee, T.H.*, +, *T-AC Mar 94* 565-568
- left-invertible generalized state space systs., disturbance rejection *Paraskevopoulos, P.N.*, +, *T-AC Jan 94* 185-190
- lin. syst. approx., time-scaling factor, Laguerre models *Wang, L.*, +, *T-AC Jul 94* 1463-1467
- rational L^1 suboptimal compensators for continuous-time systs. *Blanchini, F.*, +, *T-AC Jul 94* 1487-1492
- sampled data control systs. and tracking, fn. space approach *Yamamoto, Y.*, *T-AC Apr 94* 703-713
- semi-cancellable fraction transfer fns. in syst. theory *Bourles, H.*, *T-AC Oct 94* 2148-2153
- SISO systs., step response, undershoot *Leon de la Barra S., B.A.*, *T-AC Mar 94* 578-581
- state observer/feedback, compensators, comment *Bender, D.J.*, *T-AC Feb 94* 447-448
- Poles and zeros; cf. Zero assignment**
- Polynomial approximation**
- discrete time Hammerstein syst. ident. *Lang Zi-Qiang*, *T-AC Mar 94* 569-573
- Polynomial approximation; cf. Spline functions**
- Polynomial matrices**
- deadbeat ripple-free tracking *Jetto, L.*, *T-AC Aug 94* 1759-1764
- delta-operator formulated real polynomials, tabular method for determining root distrib. *Premaratne, K.*, +, *T-AC Feb 94* 352-355
- greatest common divisor of polynomials comput., matrix pencil based num. method *Karcanias, N.*, +, *T-AC May 94* 977-981
- J-spectral factorization *Kwakernaak, H.*, +, *T-AC Feb 94* 315-328
- lin. multivariable systs., fund. notion of equivalence *Pugh, A.C.*, +, *T-AC May 94* 1141-1145
- std. multivariable H_2 -optimal control problem, polynomial soln. *Hunt, K.J.*, +, *T-AC Jul 94* 1502-1507
- Sylvester eqns., generalised, consistency *Wimmer, H.K.*, *T-AC May 94* 1014-1016
- Polynomials**
- 1-D and 2-D digital recursive filters, stabil., Schur polynomials *Barret, M.*, +, *T-AC Nov 94* 2335-2339
- extreme-point robust stabil., discrete-time polynomials *Perez, F.*, +, *T-AC Jul 94* 1470-1472
- freq. response arcs, stable polynomial, convexity *Keqin Gu*, *T-AC Nov 94* 2262-2265
- Hurwitz and Schur polynomials, argument conditions *Bose, N.K.*, *T-AC Feb 94* 345-346
- Hurwitz polynomials props., polynomial families stabil. anal. *Duan, G.-R.*, +, *T-AC Dec 94* 2490-2494
- interval lin. control syst. robust parametric design. *Keel, L.H.*, +, *T-AC Jul 94* 1524-1530
- parametrically uncertain systs., anal. and synthesis. *Kaminsky, R.D.*, +, *T-AC Apr 94* 874-876
- polynomial differential eqns., domain of attraction estim. *Levin, A.*, *T-AC Dec 94* 2471-2475
- robust adaptive pole placement control. *Weyer, E.*, +, *T-AC Aug 94* 1665-1671
- Robust stabil. of polynomials, multilinearly depend. coeff. perturb. *Tian Yu-Ping*, +, *T-AC Mar 94* 554-558
- Routh table prop., quadratic integral eval. appl. *Beghi, A.*, +, *T-AC Dec 94* 2494-2496

- sing. root distrib. problems, displacement struct. approach. *Pal, D.*, +, *T-AC Jan 94* 238-245
- stabil. domains, assigned root location polynomials, convexity props. *Tesi, A.*, +, *T-AC Mar 94* 668-672
- stabilisation, uncertainties/numerator-denominator coupling. *Chockalingam, G.*, +, *T-AC Sep 94* 1955-1958
- strictly Hurwitz conditions, stabil. anal. appl. *Polyak, B.T.*, +, *T-AC May 94* 1147-1148
- time-delay syst. robust stabil., quasipolynomial convex directions/testing sets. *Kharitonov, V.L.*, +, *T-AC Dec 94* 2388-2397
- time-invariant lin. systs., stabil. domains inversion. *Walter, E.*, +, *T-AC Apr 94* 886-889
- uncertain syst. pole assignment, polynomial approach. *Figuerola, J.L.*, +, *T-AC Apr 94* 831-835
- Position control**
- autonomous vision-based mobile robot. *Baumgartner, E.T.*, +, *T-AC Mar 94* 493-502
- contact stabil. of simple posn. controllers, effect of time delay and discrete control. *Fiala, J.*, +, *T-AC Apr 94* 870-873
- mobile manipulator locomotion/manipulation coord. *Yamamoto, Y.*, +, *T-AC Jun 94* 1326-1332
- motion/force control, robust, with nonholonomic constraints. *Chun-Yi Su*, +, *T-AC Mar 94* 609-614
- robotic arms, repositioning control by learning. *Lucibello, P.*, *T-AC Aug 94* 1690-1694
- robot manipulators, compliant, force/posn. regulation. *Chiaverini, S.*, +, *T-AC Mar 94* 647-652
- stick-slip avoidance, PD control. *Dupont, P.E.*, *T-AC May 94* 1094-1097
- Positive real functions**
- dyn. lin. systs., pos. real lemma generalization. *Scherer, R.*, +, *T-AC Apr 94* 882-886
- pos. realness conditions, charactn. assuming controllability. *Weiss, H.*, +, *T-AC Mar 94* 540-544
- second-order dyn. systs., dissipative controller designs. *Morris, K.A.*, +, *T-AC May 94* 1056-1063
- Power demand; cf. Load forecasting**
- Power generation control, excitation**
- variable struct. generator voltage regulator design, pole assignment tech. *Aggoune, M.E.*, +, *T-AC Oct 94* 2106-2110
- Power generation control, load frequency**
- alternating renewal elec. load models. *El-Ferik, S.*, +, *T-AC Jun 94* 1184-1196
- Power generation excitation systems; cf. Power generation control, excitation**
- Power system identification**
- alternating renewal elec. load models. *El-Ferik, S.*, +, *T-AC Jun 94* 1184-1196
- Power system modeling; cf. Load modeling**
- Prediction methods**
- lin. discrete-time systs., optimum lin. recursive estim. *Carazo, A.H.*, +, *T-AC Aug 94* 1636-1638
- recursive prediction error algm. log likelihood fn. derivatives. *Hooker, M.A.*, *T-AC Mar 94* 662-664
- Predictive control**
- adaptive predictive controller, robustness anal. *Clarke, D.W.*, +, *T-AC May 94* 1052-1056
- extended horizon predictive control. *Fanyin Kong*, +, *T-AC Jul 94* 1467-1470
- SISO discrete-time systs. preview tracking. *Halpern, M.E.*, *T-AC Mar 94* 589-592
- Smith predictor for controlling proc., integrator and long dead-time. *Astrom, K.J.*, +, *T-AC Feb 94* 343-345
- stochastic adaptive prediction/MRAC. *Wei Ren*, +, *T-AC Oct 94* 2047-2060
- Probability**
- closed Jackson queueing network decentralized control. *Dye-Jyun Ma*, +, *T-AC Jul 94* 1460-1463
- distributed algs., random proc. failures. *Papavasiliopoulos, G.P.*, *T-AC May 94* 1032-1036
- integrated probabilistic data assoc. *Musicki, D.*, +, *T-AC Jun 94* 1237-1241
- mfg. systs. with prod. rate-depend. failure rates, prod. control. *Liberopoulos, G.*, +, *T-AC Apr 94* 889-895
- mfg. syst. with unreliable machines, stabil. anal. *Han-Fu Chen*, +, *T-AC Mar 94* 681-686
- multi-DOF nonlin. damping model, spectral dens. *Wei Jian Zhang*, *T-AC Feb 94* 406-410
- random param. tracking, robust algm. *Juditsky, A.*, +, *T-AC Jun 94* 1211-1221
- recursive prediction error algm. log likelihood fn. derivatives. *Hooker, M.A.*, *T-AC Mar 94* 662-664
- sens. fusion, decentralized detect. optimal thresholds. *Irving, W.W.*, +, *T-AC Apr 94* 835-838
- Production control**
- distributed prod. control methods. *Sharifnia, A.*, *T-AC Apr 94* 725-737
- failure prone mfg. syst. hedging point policies. *Jian-Qiang Hu*, +, *T-AC Sep 94* 1875-1880
- failure-prone mfg. systs., structural props. of optimal prod. controllers. *Jian-Qiang Hu*, +, *T-AC Mar 94* 640-643
- machine capacity allocation. *Nain, P.*, +, *T-AC Sep 94* 1853-1855
- mfg. systs. with prod. rate-depend. failure rates. *Liberopoulos, G.*, +, *T-AC Apr 94* 889-895
- multiple time scale approach. *Jiang, J.*, +, *T-AC Nov 94* 2292-2297
- single-machine scheduling, resource optimal control. *Cheng, T.C.E.*, +, *T-AC Jun 94* 1243-1246
- stochastic dyn. job shops/prod. planning. *Sethi, S.*, +, *T-AC Oct 94* 2061-2076
- Production control; cf. Manufacturing**
- Production systems; cf. Manufacturing scheduling**
- Proportional control**
- left-invertible generalized state space systs., disturbance rejection. *Paraskevopoulos, P.N.*, +, *T-AC Jan 94* 185-190
- robot manipulators, compliant, force/posn. regulation. *Chiaverini, S.*, +, *T-AC Mar 94* 647-652
- stick-slip avoidance, PD control. *Dupont, P.E.*, *T-AC May 94* 1094-1097
- Protocols**
- optimal multicopy Aloha. *Wong, E.W.M.*, +, *T-AC Jun 94* 1233-1236
- Pulse-width modulation**
- feedback control by multirate PWM. *Khayatian, A.*, +, *T-AC Jun 94* 1292-1297
- Pulse-width modulation, power converters**
- switched-mode power converters, multirate modeling and control design. *Khayatian, A.*, +, *T-AC Sep 94* 1848-1852
- Quantized control**
- quantized constrained systs., feedback control, appls. to neuromorphic controllers design. *Sznaier, M.*, +, *T-AC Jul 94* 1497-1502
- Queueing analysis**
- closed Jackson queueing network decentralized control. *Dye-Jyun Ma*, +, *T-AC Jul 94* 1460-1463
- DD scheduling policies, queueing networks, perform. bounds. *T-AC Aug 94* 1600-1611
- DEDS, steady state fn., nondifferentiability. *Shapiro, A.*, +, *T-AC Aug 94* 1707-1711
- deterministic/stochastic queueing networks anal. *Cheng-Shang Chang*, *T-AC May 94* 913-931
- finite queues, priority-discarding policies. *Petr, D.W.*, *T-AC May 94* 1020-1024
- FMS, FCFS scheduling policy. *Seidman, T.I.*, *T-AC Oct 94* 2166-2171
- machine capacity allocation. *Nain, P.*, +, *T-AC Sep 94* 1853-1855
- M/M/1 queue with bal. budget, optimal flow control. *Chakravorti, B.*, *T-AC Sep 94* 1918-1921
- packet switching, flow control, rate-based, monotonicity/concavity. *Budka, K.C.*, *T-AC Mar 94* 544-548
- routing, limited state inform. in queueing systs., blocking. *Sparaggis, P.D.*, +, *T-AC Jul 94* 1492-1497
- traffic models, queueing networks anal., projection techs. *Mooi Choo Chuah*, *T-AC Aug 94* 1588-1599
- Radar data processing**
- integrated probabilistic data assoc. *Musicki, D.*, +, *T-AC Jun 94* 1237-1241
- Radar signal analysis**
- integrated probabilistic data assoc. *Musicki, D.*, +, *T-AC Jun 94* 1237-1241
- range tracking loops, large deviation anal. *Dembo, A.*, +, *T-AC Feb 94* 360-364
- Radar tracking**
- integrated probabilistic data assoc. *Musicki, D.*, +, *T-AC Jun 94* 1237-1241
- Radio communication; cf. Packet radio**
- Random...; cf. Stochastic...**
- Rational functions**
- freq. response aros, stable polynomial, convexity. *Keqin Gu*, *T-AC Nov 94* 2262-2265

Realization theory; cf. Minimal realizations**Real-time systems**

constrained nonlin. syst. receding horizon control. *Shin-Yeu Lin*, *T-AC Sep 94* 1893-1899

Recursive digital filters

1-D and 2-D digital recursive filters, stabil., Schur polynomials. *Barret, M.*, +, *T-AC Nov 94* 2335-2339
optimal filtering, stochastic discrete-time systs., unknown inputs. *Borisov, A.V.*, +, *T-AC Dec 94* 2461-2464

Recursive estimation

lin. discrete-time systs., optimum lin. recursive estim. *Carazo, A.H.*, +, *T-AC Aug 94* 1636-1638
multiscale recursive estim., data fusion/regularization. *Chou, K.C.*, +, *T-AC Mar 94* 464-478
recursive prediction error algm. log likelihood fn. derivatives. *Hooker, M.A.*, *T-AC Mar 94* 662-664
stochastic lin. discrete systs., state estim. algm. *Ahmed, M.S.*, *T-AC Aug 94* 1652-1656
time-varying params., bounded error ident. *Bittanti, S.*, +, *T-AC May 94* 1106-1110

Reduced-order systems

discrete-time q-Markov cover models. *Sreeram, V.*, +, *T-AC May 94* 1102-1105
discrete-time systs., generalized q-Markov cover models. *Sreeram, V.*, *T-AC Dec 94* 2502-2505
distributed delay lin. systs., feedback stabilisation. *Feng Zheng*, +, *T-AC Aug 94* 1714-1718
full-order/low-order observers, LTI plant, unified singular syst. based theory. *Cobb, J.D.*, *T-AC Dec 94* 2497-2502
 H_∞ compensator design, min. order observers. *Hsu, C.S.*, +, *T-AC Aug 94* 1679-1681
 H_2 -optim., reduced order observer based controller. *Stoorvogel, A.A.*, +, *T-AC Feb 94* 355-360
interval systs. order reduction, Routh-Pade approx. *Bandyopadhyay, B.*, +, *T-AC Dec 94* 2454-2456
 L^2 optimal model order reduction problem, input normal form homotopy. *Ge, Y.*, +, *T-AC Jun 94* 1302-1305
model reduction, LF approx. balancing, props. *Prakash, R.*, *T-AC May 94* 1135-1141
structural wave control, reduced-order model. *Quan Wang*, +, *T-AC Aug 94* 1711-1713
unstable plants, sampled-data observers, generalized holds. *Haddad, W.M.*, +, *T-AC Jan 94* 229-234
vibr. systs. control. *Karl, W.C.*, +, *T-AC Jan 94* 222-226

Regenerative stochastic processes

DEDS, steady state fn., nondifferentiability. *Shapiro, A.*, +, *T-AC Aug 94* 1707-1711
regenerative syst. optim., infinitesimal perturb. anal. *Chong, E.K.P.*, +, *T-AC Jul 94* 1400-1410

Regulators

discrete-time H_∞ control problem, strictly proper meas. feedback. *Stoorvogel, A.A.*, +, *T-AC Sep 94* 1936-1939
finite-dimens. controllers designed for infinite-dimens. systs., state space. *Morris, K.A.*, *T-AC Oct 94* 2100-2104
mfg. scheduling syst., regulator stabilization tech. *Humes, C., Jr.*, *T-AC Jan 94* 191-196
nonlin. regulator, num. design. *Kreisselmeier, G.*, +, *T-AC Jan 94* 33-46

Relaxation methods

noncooperative equilibria computation, relax. algms. *Uryas'ev, S.*, +, *T-AC Jun 94* 1263-1267

Reliability

failure-prone mfg. systs., structural props. of optimal prod. controllers. *Jian-Qiang Liu*, +, *T-AC Mar 94* 640-643

Reliability; cf. Failure analysis; Fault tolerance**Reliability theory**

mfg. syst. with unreliable machines, stabil. anal. *Han-Fu Chen*, +, *T-AC Mar 94* 681-686

Resource management

machine capacity allocation. *Nain, P.*, +, *T-AC Sep 94* 1853-1855
single-machine scheduling, resource optimal control. *Cheng, T.C.E.*, +, *T-AC Jun 94* 1243-1246

Reviews; cf. Book reviews**Riccati equations**

four-block problem, recursive Schur-based soln. *Constantinescu, T.*, +, *T-AC Jul 94* 1476-1481
jump lin. systs., coupled Riccati eqns. *Abou-Kandil, H.*, +, *T-AC Aug 94* 1631-1636
mixed H_2/H_∞ control, Nash game approach. *Limebeer, D.J.N.*, +, *T-AC Jan 94* 69-82
model-following systs., time/freq. domain design equivalence. *Yen-Ting Hsu*, +, *T-AC Aug 94* 1722-1726

perturbed lin. systs., upper covariance bounds. *Bolzern, P.*, +, *T-AC Mar 94* 623-626

Riccati equations, algebraic

comments on "Stability margin evaluation for uncertain linear systems" (by C. Gong and S. Thompson, Jun 94 548-550). *Su, J.-H.*, *T-AC Dec 94* 2523-2524

constrained Riccati eqn. factorizations. *Weiss, M.*, *T-AC Mar 94* 677-681
continuous alg. Riccati eqn. eigenvalues. *Komaroff, N.*, *T-AC Mar 94* 532-534

discrete-time Riccati eqn., H_∞ control appl. *Stoorvogel, A.A.*, +, *T-AC Mar 94* 686-691

lin. sing. perturbed systs., Kalman filtering. *Gajic, Z.*, +, *T-AC Sep 94* 1952-1955

LTI syst. pos. real control. *Wei-qian Sun*, +, *T-AC Oct 94* 2034-2046

min. lin. plants parameterization. *Davis, L.D.*, +, *T-AC Apr 94* 849-852

multiscale systs., Kalman filters, Riccati eqns. *Chou, K.C.*, +, *T-AC Mar 94* 479-492

polynomial J-spectral factorization. *Kwakernaak, H.*, +, *T-AC Feb 94* 315-328

quadratic stabilization of continuous time systs. *Mahmoud, M.S.*, +, *T-AC Oct 94* 2135-2139

Riccati eqn., discrete alg., iter. matrix bounds. *Komaroff, N.*, *T-AC Aug 94* 1676-1678

Riccati eqn., discrete-time alg., closed loop eigenvalues. *Lin-Zhang Lu*, +, *T-AC Aug 94* 1682-1685

sampled-data robust control, exponential hold, optimal algm. *Yung-Chun Wu*, +, *T-AC Jan 94* 112-117

sing. perturbed syst. block-diagonalization, Taylor expansion. *Derbel, N.*, +, *T-AC Jul 94* 1429-1431

state delayed systs., memoryless H^∞ controllers. *Joon Hwa Lee*, +, *T-AC Jan 94* 159-162

three-block generalized/std. Riccati eqns. comparison. *Darouach, M.*, +, *T-AC Aug 94* 1755-1758

uncertain lin. systs., multivariable, stabil. margin eval. *Gong, C.*, +, *T-AC Mar 94* 548-550

Riccati equations, differential

exponential lin. quadratic optimal control, discounting. *Hopkins, W.E., Jr.*, *T-AC Jan 94* 175-178

Riccati equations, discrete-time

H_∞ control appl. *Stoorvogel, A.A.*, +, *T-AC Mar 94* 686-691

periodic discrete-time Riccati eqn., num. soln. *Hench, J.J.*, +, *T-AC Jun 94* 1197-1210

Riccati eqn., discrete alg., iter. matrix bounds. *Komaroff, N.*, *T-AC Aug 94* 1676-1678

Riccati eqn., discrete-time alg., closed loop eigenvalues. *Lin-Zhang Lu*, +, *T-AC Aug 94* 1682-1685

sampled-data robust control, exponential hold, optimal algm. *Yung-Chun Wu*, +, *T-AC Jan 94* 112-117

uncertain systs., sampled-data controller design. *Dolphus, R.M.*, *T-AC May 94* 1036-1042

Robots

adaptive control based passivity. *Yu Tang*, +, *T-AC Sep 94* 1871-1875

adaptive controller, Lyapunov stabil. *Egeland, O.*, +, *T-AC Aug 94* 1671-1673

arms, repositioning control, robust finite dimens. learning algms. *Lucibello, P.*, *T-AC Aug 94* 1690-1694

flexible joint robots, global regulation, approx. diff. *Kelly, R.*, +, *T-AC Jun 94* 1222-1224

manipulators with parametric uncertainty, robust control. *Keun-Mo Koo*, +, *T-AC Jun 94* 1230-1233

robust control, lin. estim. state feedback. *Berghuis, H.*, +, *T-AC Oct 94* 2159-2162

variable struct. control schemes. *Bin Yao*, +, *T-AC Feb 94* 371-376

Robots; cf. Manipulators...; Mobile robots...**Robots, sensing systems; cf. Robots, vision systems****Robots, vision systems**

autonomous vision-based mobile robot. *Baumgartner, E.T.*, +, *T-AC Mar 94* 493-502

Robustness

adaptive control algms., robust approach. *Gang Feng*, *T-AC Aug 94* 1738-1742

adaptive control systs. robustness. *Bodson, M.*, *T-AC Apr 94* 864-870

adaptive predictive controller, robustness anal. *Clarke, D.W.*, +, *T-AC May 94* 1052-1056

book review: New Tools for Robustness of Linear Systems (Barmish, B.R.; 1994). *T-AC Dec 94* 2525-2526

cascaded nonlin. systs., global robust stabilization. *Imura, J.-I.*, +, *T-AC May 94* 1084-1089

continuous direct adaptive control, saturation input constraint. *Cishen Zhang*, +, *T-AC Aug 94* 1718-1722

- corrections to "Robust stability and performance via fixed-order dynamic compensation: The discrete-time case" (May 93 776-782). *Haddad, W.M.*, +, *T-AC Aug 94* 1772
- discrete time-varying systs., robust adaptive controller. *Changyun Wen*, *T-AC May 94* 987-991
- D-stabil., robust, for lin. uncertain discrete delay systs. *Te-Jen Su*, +, *T-AC Feb 94* 425-428
- dual-rate control syst., sampled-data, stabil. robustness. *Tongwen Chen*, *T-AC Jan 94* 164-167
- dyn. lin. time-delay systs., robust controller design. *Mahmoud, M.S.*, +, *T-AC May 94* 995-999
- dyn. systs., robust adaptive control design. *Khorasani, K.*, *T-AC Aug 94* 1726-1732
- extreme-point robust stabil., discrete-time polynomials. *Perez, F.*, +, *T-AC Jul 94* 1470-1472
- flexible joint robots with uncertain params. and disturbances, tracking control. *Tomei, P.*, *T-AC May 94* 1067-1072
- floating platform, modeling and control. *Damen, A.A.H.*, +, *T-AC May 94* 1075-1078
- H_∞ multiobjective robust control, infinite-horizon. *Theodor, Y.*, +, *T-AC Oct 94* 2130-2134
- intersample props., robustness, and sensitivity, quantitat./qualitat. anal. *Feuer, A.*, +, *T-AC May 94* 1042-1047
- interval lin. control syst. robust parametric design. *Keel, L.H.*, +, *T-AC Jul 94* 1524-1530
- interval plants, closed-loop hyperstability. *Foa, Y.K.*, +, *T-AC Jan 94* 151-154
- lin. quadratic designs, real param. uncertainty, robust. *Douglas, J.*, +, *T-AC Jan 94* 107-111
- LQ optimal regulators, stabil. robustness. *Dohyoung Chung*, +, *T-AC Aug 94* 1698-1702
- L^2 optimal model order reduction problem, input normal form homotopy. *Ge, Y.*, +, *T-AC Jun 94* 1302-1305
- LTI discrete-time systs., robust stabil. anal. *Karan, M.*, +, *T-AC May 94* 991-995
- MIMO terminal sliding-mode control, rigid robotic manipulators. *Zhihong, M.*, +, *T-AC Dec 94* 2464-2469
- mixed H_2/H_∞ perform. objectives. *Kemin Zhou*, +, *T-AC Aug 94* 1564-1574
- model ref. control and modeling error compensation for robust perform./stabilization. *Sun, J.*, +, *T-AC Mar 94* 630-635
- motion/force control, robust, with nonholonomic constraints. *Chun-Yi Su*, +, *T-AC Mar 94* 609-614
- MRAC, robust, SISO systs. *Zhihua Qu*, +, *T-AC Nov 94* 2219-2234
- multivariable direct MRAC, identifiable parameterizations and param. convergence. *de Mathelin, M.*, +, *T-AC Aug 94* 1612-1617
- multivariable systs., quantitat. robustness measures. *Hong-Giou Chen*, +, *T-AC Apr 94* 807-810
- nonlin. servomechanism robust control. *Jie Huang*, +, *T-AC Jul 94* 1510-1513
- Nyquist envelope of interval plant family. *Hollot, C.V.*, +, *T-AC Feb 94* 391-396
- optimal cost control/filtering, uncertain lin. systs. *Petersen, I.R.*, +, *T-AC Sep 94* 1971-1977
- parametrically uncertain nonlin. feedback systs., absol. stabil. *Marquez, H.J.*, +, *T-AC Mar 94* 664-668
- parametrically uncertain nonlin. systs., robust stabilisation. *Schoenwald, D.A.*, +, *T-AC Aug 94* 1751-1755
- polynomial, strictly Hurwitz conditions, stabil. anal. appl. *Polyak, B.T.*, +, *T-AC May 94* 1147-1148
- polynomials with multilinearly depend. coeff. perturb., robust stabil. *Tian Yu-Ping*, +, *T-AC Mar 94* 554-558
- proximate time-optimal controller, robustness. *Pao, L.Y.*, +, *T-AC Sep 94* 1963-1966
- P-type learning control. *Saab, S.S.*, *T-AC Nov 94* 2298-2302
- rigid robotic manipulators, robust tracking control. *Man Zhihong*, +, *T-AC Jan 94* 154-159
- ripple free sampled-data robust servomechanism controller, exponential hold. *Yung-Chun Wu*, +, *T-AC Jun 94* 1287-1291
- robots, robust control, lin. estim. state feedback. *Berghuis, H.*, +, *T-AC Oct 94* 2159-2162
- robust adaptive regulation with min. prior knowledge, comment. *Ji Feng Zhang*, *T-AC Mar 94* 605
- robust anal., uncertainty value sets. *Eszter, E.G.*, +, *T-AC Nov 94* 2315-2318
- robust control design, cascade struct. approach. *Bonivento, C.*, +, *T-AC Apr 94* 846-849
- robust direct adaptive controllers, normalization tech. *Gang Feng*, +, *T-AC Nov 94* 2330-2334
- robust nonsingularity problem, SSV. *Ghwan-Lu Tseng*, +, *T-AC Oct 94* 2118-2122
- robust stabilization of syst., control delays. *Kajima, A.*, +, *T-AC Aug 94* 1694-1698
- robust stabil., time-varying struct. uncertainty. *Shamma, J.S.*, *T-AC Apr 94* 714-724
- robust strong stabilization via modified Popov controller synthesis. *Wang, Y.W.*, +, *T-AC Nov 94* 2284-2287
- sampled-data robust control, exponential hold, optimal algm. *Yung-Chun Wu*, +, *T-AC Jan 94* 112-117
- second-order dyn. systs., dissipative controller designs. *Morris, K.A.*, +, *T-AC May 94* 1056-1063
- single-arm dyn., robust variable struct./and switching- Σ adaptive control. *Li-Wen Chen*, +, *T-AC Aug 94* 1621-1626
- SISO systs., robust stabilization. *Olbrot, A.W.*, +, *T-AC Mar 94* 652-657
- SISO systs. with affine param. uncertainties, robust perform. *Kogan, J.*, *T-AC Jan 94* 227-229
- spectral syst. finite-dimens. approx. *Erickson, M.A.*, +, *T-AC Sep 94* 1904-1909
- μ calc., computational complexity. *Braatz, R.P.*, +, *T-AC May 94* 1000-1002
- stabil. of uncertain lin. systs., saturating actuators. *Jin-Hoon Kim*, +, *T-AC Jan 94* 202-207
- stochastic adaptive control algms., unmodified, robustness. *Radenkovic, M.S.*, +, *T-AC Feb 94* 396-400
- structurally uncertain systs., optimal Lyapunov fns. *Olas, A.*, *T-AC Jan 94* 167-171
- switching surface design for multivariable VSS. *Ju-Jang Lee*, +, *T-AC Feb 94* 414-419
- time-delay syst. robust stabil., quasipolynomial convex directions/testing sets. *Kharitonov, V.L.*, +, *T-AC Dec 94* 2388-2397
- time-invariant lin. systs., robust controller synthesis. *Rantzer, A.*, +, *T-AC Sep 94* 1802-1808
- time-variant displacement struct. and interpolation problems. *Sayed, A.H.*, +, *T-AC May 94* 960-976
- uncertain continuous-time implicit systs., regional pole placement, robustness. *Chun-Hsiung Fang*, +, *T-AC Nov 94* 2303-2307
- uncertain lin. systs., delay depend., robust stabil. *Bugong Xu*, *T-AC Nov 94* 2365
- uncertain syst. pole assignment, polynomial approach. *Figuerola, J.L.*, +, *T-AC Apr 94* 831-835
- uncertain syst. robust stabil. anal., nonquadratic Lyapunov fns. *Zelenitsovskiy, A.L.*, *T-AC Jan 94* 135-138
- uncertain syst. robust stabil., guardian map approach. *Shuoh Rern*, +, *T-AC Jan 94* 162-164
- uncertain transfer fns., value sets comp. *Gutman, P.-O.*, +, *T-AC Jun 94* 1268-1273
- VSS control design for uncertain discrete-time systs. *Wen-June Wang*, +, *T-AC Jan 94* 99-102
- worst-case ident. anal., BIBO robustness. *Partington, J.R.*, +, *T-AC Oct 94* 2171-2176
- Roots; cf. Polynomial matrices, Polynomials**
- Rotating machines; cf. Brushless rotating machines**
- Routh methods**
- delta-operator formulated real polynomials, tabular method for determining root distrib. *Premaratne, K.*, +, *T-AC Feb 94* 352-355
- Hurwitz polynomials props., polynomial families stabil. anal. *Duan, G.-R.*, +, *T-AC Dec 94* 2490-2494
- interval systs. order reduction, Routh-Pade approx. *Bandyopadhyay, B.*, +, *T-AC Dec 94* 2454-2456
- Routh table prop., quadratic integral eval. appl. *Beghi, A.*, +, *T-AC Dec 94* 2494-2496
- strictly Hurwitz conditions, stabil. anal. appl. *Polyak, B.T.*, +, *T-AC May 94* 1147-1148
- time-invariant lin. systs., stabil. domains inversion. *Waller, E.*, +, *T-AC Apr 94* 886-889
- Sample-and-hold circuits**
- intersample props., robustness, and sensitivity, quantitat./qualitat. anal. *Feuer, A.*, +, *T-AC May 94* 1042-1047
- Sampled-data systems; cf. Discrete-time systems**
- Sampling methods**
- stochastic lin. syst. param. hypothesis testing, stat. sampling. *Duncan, T.E.*, +, *T-AC Jan 94* 118-122
- Scheduling**
- FMS, FCFS scheduling policy. *Seidman, T.I.*, *T-AC Oct 94* 2166-2171
- machine capacity allocation. *Nain, P.*, +, *T-AC Sep 94* 1853-1855
- queueing networks and scheduling policies, perform. bounds. *T-AC Aug 94* 1600-1611

Scheduling; cf. Manufacturing scheduling; Production control

Search methods

global optim., adaptive partitioned random search. *Bo Tang, Z.*, *T-AC Nov 94* 2235-2244

Sensitivity

biased/unbiased controllers LTR design. *Turan, L.*, +, *T-AC Mar 94* 601-605

continuous time nonminimum phase lin. systs., gain margin improvement. *Wei-Yong Yan, +*, *T-AC Nov 94* 2347-2354

intersample props., robustness, and sensitivity, quantitat./qualitat. anal. *Feuer, A.*, +, *T-AC May 94* 1042-1047

Nyquist envelope of interval plant family *Hollot, C V.*, +, *T-AC Feb 94* 391-396

observer design, eigenvalue assignment, residual generation. *Magni, J-F.*, +, *T-AC Feb 94* 441-447

SISO-distributed plants, optimal mixed sensitivity. *Flamm, D S.*, +, *T-AC Jun 94* 1150-1165

Sensitivity; cf. Robustness

Sensors; cf. Multisensor systems

Servosystems

H_{∞} -control, state availability rel., pole/zero cancellations *Miyamoto, S.*, +, *T-AC Feb 94* 379-381

nonlin. servomechanism robust control. *Jie Huang, +*, *T-AC Jul 94* 1510-1513

ripple free sampled-data robust servomechanism controller, exponential hold *Yung-Chun Wu, +*, *T-AC Jun 94* 1287-1291

sampled-data robust control, exponential hold, optimal algm. *Yung-Chun Wu, +*, *T-AC Jan 94* 112-117

uncertain SISO min phase lin syst., nonlin universal servomechanism. *Ryan, E P.*, *T-AC Apr 94* 753-761

Set theory

constrained continuous-time systs with cone props., pos. invariant sets. *Tarbouriech, S.*, +, *T-AC Feb 94* 401-405

corrections to "Positively invariant sets for constrained continuous-time systems with cone properties" (Feb 94 401-405). *Tarbouriech, S.*, *T-AC Aug 94* 1771

discrete-time lin systs., pos. invariant sets. *De Santis, E.*, *T-AC Jan 94* 245-249

discrete-time uncertain systs., ultimate boundedness control, Lyapunov *Blarchini, F.*, *T-AC Feb 94* 428-433

matching principle for systs., restricted inputs *Rutland, N.K.*, *T-AC Mar 94* 550-553

output regulation, necessary condition. *Lucibello, P.*, *T-AC Mar 94* 558-559

robust anal., uncertainty value sets *Eszter, E.G.*, +, *T-AC Nov 94* 2315-2318

time-invariant lin systs., stabil. domains inversion. *Walter, E.*, +, *T-AC Apr 94* 886-889

Signal analysis

multiscale recursive estim., data fusion/regularization. *Chou, K.C.*, +, *T-AC Mar 94* 464-478

Signal analysis; cf. Radar signal analysis

Signal detection

sens fusion, decentralized detect optimal thresholds. *Irving, W.W.*, +, *T-AC Apr 94* 835-838

Signal estimation; cf. Estimation

Signal processing

recursive ident., nonlin. Wiener model *Wigren, T.*, *T-AC Nov 94* 2191-2206

Signal processing; cf. Array processing; Estimation; Filtering; Image processing

Signal representations

multiscale recursive estim., data fusion/regularization. *Chou, K.C.*, +, *T-AC Mar 94* 464-478

multiscale systs., Kalman filters, Riccati eqns. *Chou, K.C.*, +, *T-AC Mar 94* 479-492

Signal resolution

multiscale recursive estim., data fusion/regularization. *Chou, K.C.*, +, *T-AC Mar 94* 464-478

Signal sampling/reconstruction; cf. Sample-and-hold circuits

Singularly perturbed systems

discrete-time q-Markov cover models. *Sreeram, V.*, +, *T-AC May 94* 1102-1105

H^{∞} -optimal control for sing. perturbed systs., imperfect state meas. *Zigang Pan, +*, *T-AC Feb 94* 280-299

lin. sing. perturbed systs., Kalman filtering. *Gajic, Z.*, +, *T-AC Sep 94* 1952-1955

sing. perturbed control systs. and nonlin. differential-alg. eqns. *Krishnan, H.*, +, *T-AC May 94* 1079-1084

sing. perturbed syst. block-diagonalization, Taylor expansion. *Derbel, N.*, +, *T-AC Jul 94* 1429-1431

Singular systems

closed-loop vibr. control, Youla parameterization. *Kabamba, P.T.*, +, *T-AC Jul 94* 1455-1459

continuous descriptor systs., adaptive observer. *Uetake, Y.*, *T-AC Oct 94* 2095-2100

descriptor systs., lin. feedback closed-loop Stackelberg strategy. *Hua Xu, +*, *T-AC May 94* 1097-1102

full-order/low-order observers, LTI plant, unified singular syst. based theory. *Cobb, J.D.*, *T-AC Dec 94* 2497-2502

robust adaptive pole placement control. *Weyer, E.*, +, *T-AC Aug 94* 1665-1671

time-invariant lin. syst. adaptive control. *Karason, S P.*, +, *T-AC Nov 94* 2325-2330

Sliding-mode control; cf. Variable-structure systems

Smoothing methods

learning, inverse pass., I/O map, update-spline-smoothing. *Heiss, M.*, *T-AC Feb 94* 259-268

Software packages

complex discrete-time control systs., simul. automation. *Ellis, R.D.*, +, *T-AC Sep 94* 1795-1801

Space vehicle control

rigid spacecraft adaptive attitude control *Egeland, O.*, +, *T-AC Apr 94* 842-846

rigid spacecraft systs controllability, central gravitational field. *Lian, K-Y.*, +, *T-AC Dec 94* 2426-2441

Space vehicle tracking

rigid spacecraft adaptive attitude control. *Egeland, O.*, +, *T-AC Apr 94* 842-846

Sparse matrices

large sparse Lyapunov eqns., approx. soln. *Gudmundsson, T.*, +, *T-AC May 94* 1110-1114

Spectral factorization

constrained Riccati eqn factorizations *Weiss, M.*, *T-AC Mar 94* 677-681

polynomial J-spectral factorization. *Kwakernaak, H.*, +, *T-AC Feb 94* 315-328

Spline functions

learning, inverse pass., I/O map, update-spline-smoothing. *Heiss, M.*, *T-AC Feb 94* 259-268

Stability

1-D and 2-D digital recursive filters, stabil., Schur polynomials. *Barret, M.*, +, *T-AC Nov 94* 2335-2339

adaptive control by multiple models/switching. *Narendra, K.S.*, +, *T-AC Sep 94* 1861-1866

adaptive control systs., transient bounds *Zhuquan Zang, +*, *T-AC Jan 94* 171-175

adaptive pole placement without excitation probing sigs. *Lozano, R.*, +, *T-AC Jan 94* 47-58

adaptive stabilization, nonlin. time-varying controller *Miller, D E.*, *T-AC Jul 94* 1347-1359

almost sure sample stabil. of nonlin. stochastic dyn. systs *Zhi Yu Zhang, +*, *T-AC Mar 94* 560-565

asynchronous sampled data syst. control, design *Voulgaris, P.*, *T-AC Jul 94* 1451-1455

asynchronous systs. with Poisson transits., stabil. *Leland, R P.*, *T-AC Jan 94* 182-185

bilinear systs., adaptive stabilization, least sq. *Xi Sun, +*, *T-AC Jan 94* 207-211

CARMA plants, stabilizing I-O receding horizon control. *Chisci, L.*, +, *T-AC Mar 94* 614-618

closed-loop vibr. control, Youla parameterization. *Kabamba, P.T.*, +, *T-AC Jul 94* 1455-1459

comments on "Stability margin evaluation for uncertain linear systems" (by C. Gong and S. Thompson, Jun 94 548-550). *Su, J-H.*, *T-AC Dec 94* 2523-2524

constrained nonlin. syst receding horizon control. *Shin-Yeu Lin, T-AC Sep 94* 1893-1899

constrained robots, force/motion control. *Grabbe, M.T.*, +, *T-AC Jan 94* 179

contact stabil. of simple posn. controllers, effect of time delay and discrete control. *Fiala, J.*, +, *T-AC Apr 94* 870-873

continuous control syst., digital equiv. design methods comparison. *Hall, S.R.*, *T-AC Feb 94* 420-421

continuous descriptor systs., adaptive observer. *Uetake, Y.*, *T-AC Oct 94* 2095-2100

continuous-time multivariable syst. reduction *Krajewski, W.*, +, *T-AC Oct 94* 2126-2129

corrections to "On the stability proof of adaptive schemes with static normalizing signal and parameter projection" (Jan 93 170-173). *Ioannou, P.*, *T-AC Apr 94* 896

cyclicly switched param.-adaptive control systs., MIMO design models/internal regulators. *Morse, A.S.*, +, *T-AC Sep 94* 1809-1818

- delta-operator formulated real polynomials, tabular method for determining root distrib. Premaratne, K., +, T-AC Feb 94 352-355
- descriptor systs. control by output feedback. Lovass-Nagy, V., +, T-AC Jul 94 1507-1509
- deterministic/stochastic queuing networks anal. Cheng-Shang Chang, T-AC May 94 913-931
- diagonal decoupling, dyn. output feedback/const. precompensator. Eldem, V., T-AC Mar 94 503-511
- discrete-time adaptive nonlin. syst., least sq. estimator. Kanellakopoulos, I., T-AC Nov 94 2362-2365
- discrete-time filters from high-order s-to-z mappings. Schneider, A.M., +, T-AC Feb 94 435-441
- discrete-time nonlin. syst. control, stabil. property. Jie Huang, +, T-AC Nov 94 2307-2311
- discrete time nonlin. systs., stabil., time delayed feedback. Xueshan Yang, +, T-AC Mar 94 585-589
- discrete-time q-Markov cover models. Sreeram, V., +, T-AC May 94 1102-1105
- discrete-time Riccati eqn., H_∞ control appl. Stoorvogel, A.A., +, T-AC Mar 94 686-691
- dissipative H_2/H_∞ controller synthesis. Haddad, W.M., +, T-AC Apr 94 827-831
- distributed delay lin. systs., feedback stabilisation. Feng Zheng, +, T-AC Aug 94 1714-1718
- D-stabil., robust, for lin. uncertain discrete delay systs. Te-Jen Su, +, T-AC Feb 94 425-428
- evol. eqns., parab. partial DE, exponential stabil. Keum-Shik Hong, +, T-AC Jul 94 1432-1436
- finite-dimens. controllers designed for infinite-dimens. systs., state space Morris, K.A., T-AC Oct 94 2100-2104
- freq. response arcs, stable polynomial, convexity. Keqin Gu, T-AC Nov 94 2262-2265
- fuzzy control systs., supervisory controller. Li-Xiu Wang, T-AC Sep 94 1845-1847
- interconnected systs., decentralized adaptive regulation Changyun Wen, T-AC Oct 94 2163-2166
- internally stable nonlin. systs. with disturbances, parameterization. Hammer, J., T-AC Feb 94 300-314
- interval LTI systs., stabilization. Kehui Wei, T-AC Jan 94 22-32
- interval matrices, necessary and sufficient conditions for Hurwitz/Schur stabil. Kaiming Wang, +, T-AC Jun 94 1251-1255
- interval matrices stabil. conditions. Sezer, M.E., +, T-AC Feb 94 368-371
- jump-Markov systs., stabilizing control law. Dufour, F., +, T-AC Nov 94 2354-2357
- large scale interconnected systs. with delays, decentralized stabilization Zhongzhi Hu, T-AC Jan 94 180-182
- lin. quadratic control, convex prog., num. method. Peres, P.L.D., +, T-AC Jan 94 198-202
- lin. uncertain syst. robust stabil., guardian map approach. Shuoh Rern, +, T-AC Jan 94 162-164
- LTI syst. pos. real control. Weiqian Sun, +, T-AC Oct 94 2034-2046
- Lyapunov stabil. theory of nonsmooth systs. Shevitz, D., +, T-AC Sep 94 1910-1914
- manipulators with parametric uncertainty, robust control. Keun-Mo Koo, +, T-AC Jun 94 1230-1233
- mfg. scheduling syst., regulator stabilization tech. Humes, C., Jr., T-AC Jan 94 191-196
- mfg. syst. with unreliable machines, stabil. anal. Han-Fu Chen, +, T-AC Mar 94 681-686
- MIMO lin. systs. with sens./actuator failures, stabil. Gundes, A.N., T-AC Jun 94 1224-1230
- min.-phase lin. plants, stabilizing controllers parameterization. Glaria, J.J., +, T-AC Feb 94 433-434
- mixed H_2/H_∞ control, Nash game approach. Limebeer, D.J.N., +, T-AC Jan 94 69-82
- multirate sampled-data systs. struct. props. Longhi, S., T-AC Mar 94 692-696
- multivariable nonlin. controller, vibr. damping. Kanestrom, R.K., +, T-AC Sep 94 1925-1928
- multivariable syst. decentralized control, closed-loop props. Campo, P.J., +, T-AC May 94 932-943
- neutral systs., stabil. and stabilizability. Logemann, H., +, T-AC Jan 94 138-143
- nonlin. feedback parametrically uncertain systs., absol. stabil. Marquez, H.J., +, T-AC Mar 94 664-668
- nonlin. servomechanism robust control. Jie Huang, +, T-AC Jul 94 1510-1513
- nonlin. systs., triangular struct., adaptive control. Seto, D., +, T-AC Jul 94 1411-1428
- nonlin. uncertain systs. tracking control. Song, Y.D., +, T-AC Sep 94 1866-1871
- nonnegative feedback control, oscill. stabilisation. Zaslavsky, B., T-AC Jun 94 1273-1276
- output feedback problem in lin. systs., bilinear formulation. Syrmos, V.L., +, T-AC Feb 94 410-414
- output feedback stabilization of nonlin. systs. Tsirlas, J., +, T-AC Apr 94 806
- output regulation, necessary condition. Lucibello, P., T-AC Mar 94 558-559
- parab. systs. direct adaptive control. Keum Shik Hong, +, T-AC Oct 94 2018-2033
- param.-adaptive control, cyclic switching strategy. Pait, F.M., +, T-AC Jun 94 1172-1183
- partially obs. discrete-time nonlin. systs., risk-sensitive control and dyn. games. James, M.R., +, T-AC Apr 94 780-792
- periodically time-varying lin. discrete-time plants, decentralized control. Khargonekar, P.P., +, T-AC Apr 94 877-882
- pos. real systs. nonlin. controllers. Bernstein, D.S., +, T-AC Jul 94 1513-1517
- prod. control methods, distributed, stabil. and perform. Sharifnia, A., T-AC Apr 94 725-737
- quadratic stabilization of continuous time systs. Mahmoud, M.S., +, T-AC Oct 94 2135-2139
- rational L^1 suboptimal compensators for continuous-time systs. Blanchini, F., +, T-AC Jul 94 1487-1492
- rational transfer fn., optimal L_∞ approx. Kavranoğlu, D., +, T-AC Sep 94 1899-1904
- robot adaptive control based on passivity. Yu Tang, +, T-AC Sep 94 1871-1875
- robot controller, adaptive, Lyapunov stabil. Egeland, O., +, T-AC Aug 94 1671-1673
- robots, trajectory stabilization for systs., nonholonomic constraints Walsh, G., +, T-AC Jan 94 216-222
- robust adaptive pole placement control. Weyer, E., +, T-AC Aug 94 1665-1671
- row-by-row stable decoupling, static state feedback, struct. soln. Martinez Garcia, J.C., +, T-AC Dec 94 2457-2460
- sampled-data systs., wordlength constraint, stabil./perform. Fialho, I.J., +, T-AC Dec 94 2476-2481
- SISO LTI discrete-time systs., param.-adaptive controller. Kreisselmeier, G., T-AC Sep 94 1819-1826
- stabilization of discrete-time nonlin. systs., global Byrnes, C.I., +, T-AC Jan 94 83-98
- stabil. of uncertain lin. systs., saturating actuators. Jin-Hoon Kim, +, T-AC Jan 94 202-207
- static output feedback for stabilization. Pimpalkhare, A.A., +, T-AC May 94 1148
- stochastic adaptive control algms., unmodified, robustness Radenkovic, M.S., +, T-AC Feb 94 396-400
- stochastic adaptive prediction/MRAC Wei Ren, +, T-AC Oct 94 2047-2060
- structurally uncertain systs., optimal Lyapunov fns. Olas, A., T-AC Jan 94 167-171
- switching surface design for multivariable VSS Ju-Jang Lee, +, T-AC Feb 94 414-419
- three-block generalized/std Riccati eqns. comparison Darouach, M., +, T-AC Aug 94 1755-1758
- time-invariant lin. syst. adaptive control Karason, S.P., +, T-AC Nov 94 2325-2330
- time-invariant lin. systs., stabil. domains inversion. Walter, E., +, T-AC Apr 94 886-889
- time lag nonlin. systs., vibr. control. Lehman, B., +, T-AC May 94 898-912
- time-varying functional diff. eqns., delay independent stabil. conditions/decay estim. Lehman, B., +, T-AC Aug 94 1673-1676
- time varying systs., discrete time adaptive controller, global stabil. Radenkovic, M.S., +, T-AC Nov 94 2357-2361
- tip mass/cable syst., stabilization. Morgul, O., +, T-AC Oct 94 2140-2145
- transfer matrices preconditioning. Rotstein, B., T-AC Nov 94 2287-2292
- uncertain discrete-time systs., variable struct. control design Myszkowski, P., +, T-AC Nov 94 2366-2367
- uncertain lin. systs., multivariable, stabil. margin eval. Gong, C., +, T-AC Mar 94 548-550
- uncertain SISO min. phase lin. syst., nonlin. universal servomechanism Ryan, E.P., T-AC Apr 94 753-761
- uncertain syst. robust stabil. anal., nonquadratic Lyapunov fns. Zelenitsovskiy, A.L., T-AC Jan 94 135-138
- uncertain systs., sampled-data controller design Dolphus, R.M., T-AC May 94 1036-1042
- uncertain systs. stabilizing control design, quasiconvex optim. Keqin Gu, T-AC Jan 94 127-131
- variable struct. MRAC, I/O based, anal./design. Liu Hsu, +, T-AC Jan 94 4-21
- vibr. systs. control. Karl, W.C., +, T-AC Jan 94 222-226

Stability; cf. Absolute stability, Asymptotic stability, Input-output stability, Lyapunov methods, Nyquist stability, Robustness, Routh methods

State estimation

- biased/unbiased controllers LTR design *Turan L* + *T-AC Mar 94* 601-605
- discrete-time Markovian jump lin systs, lin min MSE estim *Costa O L V*, *T-AC Aug 94* 1685-1689
- disturbance decoupled observer design *Hou M* + *T-AC Jun 94* 1338-1341
- full-order/low-order observers, I FI plant, unified singular syst based theory *Cobb J D*, *T-AC Dec 94* 2497-2502
- guaranteed estim problem, Kalman-Bucy filter appl *Golovan A* + *T-AC Jun 94* 1282-1286
- lin discrete-time systs, optimum lin recursive estim *Carazo A H* + *T-AC Aug 94* 1636-1638
- MIMO stochastic discrete syst state estim *Liu Danyang* + *T-AC Oct 94* 2087-2091
- nonlin syst state estim, conditionally min algm *Pankov A R* + *T-AC Aug 94* 1617-1620
- optimal cost control/filtering, uncertain lin systs *Petersen I R* + *T-AC Sep 94* 1971-1977
- optimal filtering, stochastic discrete-time systs, unknown inputs *Borisov A V* + *T-AC Dec 94* 2461-2464
- risk-sensitive estim and differential game *Banavar R N* + *T-AC Sep 94* 1914-1918
- stochastic lin discrete systs, state estim algm *Ahmed M S* *T-AC Aug 94* 1652-1656

State estimation; cf. Kalman filtering Observers

State feedback

- adaptive tracking, feedback lin systs *Marino R* + *T-AC Jun 94* 1314-1319
- CARMA plants stabilizing I-O receding horizon control *Chisci I* + *T-AC Mar 94* 614-618
- cascaded nonlin systs global robust stabilization *Imura J I* + *T-AC May 94* 1084-1089
- certainly equivalence control dyn games *James M R* *T-AC Nov 94* 2321-2324
- constrained continuous-time systs with cone props pos invariant sets *Tarbouriech S* + *T-AC Feb 94* 401-405
- corrections to 'Positively invariant sets for constrained continuous time systems with cone properties' (1 Feb 94 401-405) *Tarbouriech S* *T-AC Aug 94* 1771
- discrete-time nonlin control systs design via smooth feedback *Wei Lin* + *T-AC Nov 94* 2340-2346
- disturbance decoupling constrained Sylvester eqns *Syrmos V L* *T-AC Apr 94* 797-803
- failure detect /isolation/accommodation syst *Chia Chi Tsui* *T-AC Nov 94* 2318-2321
- feedback control by multirate PWM *Khayatian A* + *T-AC Jun 94* 1292-1297
- H_∞ control of discrete-time uncertain systs *Grolmel J C* + *T-AC May 94* 1072-1075
- H_∞ -control state availability rel pole/zero cancellations *Miyamoto S* + *T-AC Feb 94* 379-381
- H_∞ -optimal reduced order observer based controller *Stoorvogel A A* + *T-AC Feb 94* 355-360
- H^∞ -optimal control for sing perturbed systs imperfect state meas *Zigang Pan* + *T-AC Feb 94* 280-299
- input-output decoupling nonlin interactor appl *Di Benedetto M D* + *T-AC Jun 94* 1246-1250
- input-output lin state equivalence and decoupling *In-Joong Ha* + *T-AC Nov 94* 2269-2274
- interval LFI systs, stabilization *Kehui Wei* *T-AC Jan 94* 22-32
- left-invertible generalized state space systs, disturbance rejection *Paraskevopoulos P N* + *T-AC Jan 94* 185-190
- MIMO nonlin syst I/O pseudolinarization *Lawrence D A* + *T-AC Nov 94* 2207-2218
- mixed H_2/H_∞ control, Nash game approach *Limebeer D J N* + *T-AC Jan 94* 69-82
- optimal cost control/filtering, uncertain lin systs *Petersen I R* + *T-AC Sep 94* 1971-1977
- robots, robust control lin estim state feedback *Berghuis H* + *T-AC Oct 94* 2159-2162
- row-by-row stable decoupling, static state feedback, struct soln *Martinez Garcia J C* + *T-AC Dec 94* 2457-2460
- stabilization of discrete-time nonlin systs, feedback equivalence *Byrnes C I* + *T-AC Jan 94* 83-98
- state deadbeat control problem general soln *Eldem Y* + *T-AC May 94* 1002-1006
- state observer/feedback compensators, comment *Bender D J* *T-AC Feb 94* 447-448

uncertain nonlin systs, struct invariance *Castro-Linares, R*, +, *T-AC Oct 94* 2154-2158

vector discrete-event syst controller synthesis *Yong Li* + *T-AC Mar 94* 512-531

State-space methods

- 2D general discrete state-space models eigenvalues calc *Zou Yun* + *T-AC Jul 94* 1436-1439
 - bilinear systs, state space anal, orthogonal series approach *Paraskevopoulos P N* +, *T-AC Apr 94* 793-797
 - block multirate input-output model for sampled-data control systs *Jakubowski A M* + *T-AC May 94* 1145-1147
 - comments on "Stability margin evaluation for uncertain linear systems" (by C. Gong and S. Thompson, Jun 94 548-550) *Su J-H* *T-AC Dec 94* 2523-2524
 - continuous descriptor systs, adaptive observer *Uetake Y* *T-AC Oct 94* 2095-2100
 - dyn shock-error models, online estim *Krishnamurthy V* *T-AC May 94* 1129-1135
 - extended Chandrasekhar recursions *Sayed A H* + *T-AC Mar 94* 619-623
 - fast ident of state-space models *Young Man Cho* + *T-AC Oct 94* 2004-2017
 - finite-dimens controllers designed for infinite-dimens systs, state space *Morris K A* *T-AC Oct 94* 2100-2104
 - four-block problem, recursive Schur-based soln *Constantinescu T* + *T-AC Jul 94* 1476-1481
 - generalized Chandrasekhar recursions from generalized Schur algm *Sayed A H* + *T-AC Nov 94* 2265-2269
 - H_∞ -control, state availability rel, pole/zero cancellations *Miyamoto S* + *T-AC Feb 94* 379-381
 - H_∞ optim, time-domain constraints *Rotstein H* + *T-AC Apr 94* 762-779
 - ident, state-space freq domain approach *Bayard D S* *T-AC Sep 94* 1880-1885
 - left-invertible generalized state space systs disturbance rejection *Paraskevopoulos P N* + *T-AC Jan 94* 185-190
 - lin-quadratic zero-sum differential games for generalized state space systs *Hua Xu* + *T-AC Jan 94* 143-147
 - lin uncertain syst robust stabil guardian map approach *Shuoh Rern* + *T-AC Jan 94* 162-164
 - I FI syst pos real control *Wei-qian Sun* + *T-AC Oct 94* 2034-2046
 - mixed H_2/H_∞ perform objectives *Kemin Zhou* + *T-AC Aug 94* 1564-1574
 - mixed H_2/H_∞ perform objectives optimal control *Doyle J* + *T-AC Aug 94* 1575-1587
 - multichannel gain margin improvement, sampled-data hold fns *Chang Yang* + *T-AC Mar 94* 657-661
 - multiscale recursive estim data fusion/regularization *Chou K C* + *T-AC Mar 94* 464-478
 - multiscale systs Kalman filters Riccati eqns *Chou K C* + *T-AC Mar 94* 479-492
 - nonlin syst H_∞ control output-feedback based *Lu W-M* + *T-AC Dec 94* 2517-2524
 - output regulation, necessary condition *Lucibello P* *T-AC Mar 94* 558-559
 - pos realness conditions charactn assuming controllability *Weiss H* + *T-AC Mar 94* 540-544
 - rational transfer fn, optimal l_∞ approx *Kavranoglu D* + *T-AC Sep 94* 1899-1904
 - recursive prediction error algm log likelihood fn derivatives *Hooker M A* *T-AC Mar 94* 662-664
 - uncertain lin systs, multivariable, stabil margin eval *Gong C* + *T-AC Mar 94* 548-550
- # Statistics
- alternating renewal elec load models *El-Ferik S* + *T-AC Jun 94* 1184-1196
 - discrete time Hammerstein syst ident *Lang Zi-Qiang* *T-AC Mar 94* 569-573
 - dyn syst simul, fast parallel recursive aggregation *Tsai W K* + *T-AC Mar 94* 534-540
 - guaranteed param estim problem with uncertain stats, Kalman-Bucy filter accuracy *Matasov A I* *T-AC Mar 94* 635-639
 - Kautz models for syst ident *Wahlberg B* *T-AC Jun 94* 1276-1282
 - least sq estim in white noise, convergence *Nassiri-Toussi K* + *T-AC Feb 94* 364-368
 - MIMO stochastic discrete syst state estim *Liu Danyang* + *T-AC Oct 94* 2087-2091
 - model struct selection test, instrumental variable, statist props *Hoar Nghia Duong* + *T-AC Jan 94* 211-215
 - multi-DOF nonlin damping model, spectral dens *Weijian Zhang* *T-AC Feb 94* 406-410
 - multi-point boundary value stochastic systs, stationarity/reciprocity *Jie Chen* + *T-AC May 94* 1114-1116

- stochastic lin. discrete systs., state estim. algm. *Ahmed, M.S.*, *T-AC Aug 94* 1652-1656
- Stochastic approximation**
 2D stochastic approx., convergence/diff. eqn. limit. *Dye-Jyun Ma*, + , *T-AC Jul 94* 1439-1442
 recursive ident., nonlin. Wiener model. *Wigren, T.*, *T-AC Nov 94* 2191-2206
- Stochastic differential equations**
 almost sure sample stabil. of nonlin. stochastic dyn. systs. *Zhi Yu Zhang*, + , *T-AC Mar 94* 560-565
 infinite horizon optimal control of stochastic systs. *Runolfsson, T.*, *T-AC Aug 94* 1551-1563
- Stochastic games**
 flow control, zero sum Markov games. *Altman, E.*, *T-AC Apr 94* 814-818
 infinite horizon optimal control of stochastic systs. *Runolfsson, T.*, *T-AC Aug 94* 1551-1563
 risk-averse decentralized discrete-time LEQG games. *Srikant, R.*, *T-AC Apr 94* 861-864
- Stochastic optimal control**
 infinite horizon optimal control of stochastic systs. *Runolfsson, T.*, *T-AC Aug 94* 1551-1563
 lin. exponential Gaussian control problem. *Chih-Hai Fan*, + , *T-AC Oct 94* 1986-2003
 partially obs. discrete-time nonlin. systs., risk-sensitive control and dyn. games. *James, M.R.*, + , *T-AC Apr 94* 780-792
- Stochastic optimal control, linear systems; cf. Linear-quadratic-Gaussian control**
- Stochastic processes**
 distributed algs., random proc. failures. *Papavasiliopoulos, G.P.*, *T-AC May 94* 1032-1036
 dyn. job shops/prod. planning. *Sethi, S.*, + , *T-AC Oct 94* 2061-2076
 event graphs, stochastic, cycle time. *Proth, J.-M.*, + , *T-AC Jul 94* 1482-1486
 machine capacity allocation. *Nain, P.*, + , *T-AC Sep 94* 1853-1855
 multiscale recursive estim., data fusion/regularization. *Chou, K.C.*, + , *T-AC Mar 94* 464-478
 packet switching, flow control, rate-based, monotonicity/concavity. *Budka, K.C.*, *T-AC Mar 94* 544-548
 routing, limited state inform. in queueing systs., blocking. *Sparaggis, P.D.*, + , *T-AC Jul 94* 1492-1497
 timed-event stochastic graphs, superposn. props./perform. bounds. *Xiao-Lan Xie*, *T-AC Jul 94* 1376-1386
- Stochastic processes; cf. Autoregressive moving-average processes; Innovations methods (stochastic processes); Laguerre processes; Markov processes; Poisson processes; Regenerative stochastic processes; Time-varying stochastic processes; Wiener processes**
- Stochastic systems**
 adaptive stochastic control algs., unmodified, robustness. *Radenkovic, M.S.*, + , *T-AC Feb 94* 396-400
 adaptive stochastic prediction/MRAC. *Wei Ren*, + , *T-AC Oct 94* 2047-2060
 almost sure sample stabil. of nonlin. stochastic dyn. systs. *Zhi Yu Zhang*, + , *T-AC Mar 94* 560-565
 bilinear systs., adaptive stabilization, least sq. *Xi Sun*, + , *T-AC Jan 94* 207-211
 exact lin. of stochastic dyn. systs. *Socha, L.*, *T-AC Sep 94* 1980-1984
 Gaussian stochastic control syst. tuning. *van Schuppen, J.H.*, *T-AC Nov 94* 2178-2190
 lin. discrete-time systs., state estim. *Ahmed, M.S.*, *T-AC Aug 94* 1652-1656
 MIMO stochastic discrete syst. state estim. *Liu Danyang*, + , *T-AC Oct 94* 2087-2091
 min. stochastic realizations, parameterization. *Ferrante, A.*, *T-AC Oct 94* 2122-2126
 multipoint boundary value stochastic systs., stationarity/reciprocity. *Jie Chen*, + , *T-AC May 94* 1114-1116
 nonlin. syst. state estim., conditionally min. algm. *Pankov, A.R.*, + , *T-AC Aug 94* 1617-1620
 optimal filtering, stochastic discrete-time systs., unknown inputs. *Borisov, A.V.*, + , *T-AC Dec 94* 2461-2464
 param. hypothesis testing, stochastic lin. systs., stat. sampling. *Duncan, T.E.*, + , *T-AC Jan 94* 118-122
 risk-averse decentralized discrete-time LEQG games. *Srikant, R.*, *T-AC Apr 94* 861-864
- Stochastic systems; cf. Stochastic optimal control**
- Suboptimal control**
 cat. falling, near-optimal nonholonomic motion planning, appl. *Fernandes, C.*, + , *T-AC Mar 94* 450-463
 model matching, suboptimal perfect, with noise, MRACS. *Muloh, Y.*, + , *T-AC Feb 94* 422-425
 rational L^1 suboptimal compensators for continuous-time systs. *Blanchini, F.*, + , *T-AC Jul 94* 1487-1492
- Switched systems**
 common Lyapunov fn., stable LTI systs., commuting A-matrices. *Narendra, K.S.*, + , *T-AC Dec 94* 2469-2471
- Synchronization**
 timed discrete-event systs., supervisory control. *Brandin, B.A.*, + , *T-AC Feb 94* 329-342
- Synchronous generator excitation; cf. Power generation control, excitation**
- System identification**
 cyclicly switched param.-adaptive control systs., MIMO design models/internal regulators. *Morse, A.S.*, + , *T-AC Sep 94* 1809-1818
 distributed param. systs., adaptive estim., persistence of excitation. *Demetriou, M.A.*, + , *T-AC May 94* 1117-1123
 finite-dimens. model validation, output error, test horizon. *Hoai Nghia Duong*, + , *T-AC Jan 94* 102-106
 floating platform, modeling and control. *Damen, A.A.H.*, + , *T-AC May 94* 1075-1078
 H^∞ control syst. ident., least sq. methods, comment. *Livstone, M.M.*, + , *T-AC Jul 94* 1531
 Kautz models for syst. ident. *Wahlberg, B.*, *T-AC Jun 94* 1276-1282
 least sq. estim., efficient algm. *Rafajlowicz, E.*, + , *T-AC Jun 94* 1241-1243
 norms for syst. ident. *Massoumnia, M.-A.*, + , *T-AC May 94* 1027-1031
 param.-adaptive control, cyclic switching strategy. *Pait, F.M.*, + , *T-AC Jun 94* 1172-1183
 recursive ident., nonlin. Wiener model. *Wigren, T.*, *T-AC Nov 94* 2191-2206
 robust adaptive pole placement control. *Weyer, E.*, + , *T-AC Aug 94* 1665-1671
 slowly varying and LTI systs., ident. and uncertainty principles. *Zames, G.*, + , *T-AC Sep 94* 1827-1838
 state-space freq. domain method. *Bayard, D.S.*, *T-AC Sep 94* 1880-1885
 state-space models fast ident. *Young Man Cho*, + , *T-AC Oct 94* 2004-2017
 struct. selection test, instrumental variable, statist. props. *Hoai Nghia Duong*, + , *T-AC Jan 94* 211-215
 time-varying params., bounded error ident. *Bitanti, S.*, + , *T-AC May 94* 1106-1110
 Wiener systs. nonparametric ident. *Greblicki, W.*, *T-AC Oct 94* 2077-2086
 worst-case ident. anal., BIBO robustness. *Partington, J.R.*, + , *T-AC Oct 94* 2171-2176
 worst case ident., H^∞ model validation. *Guoxiang Gu*, *T-AC Aug 94* 1657-1661
 worst-case syst. ident., time complexity. *Poolla, K.*, + , *T-AC May 94* 944-950
- System identification; cf. Parameter identification, Power system identification**
- System reliability; cf. Reliability**
- Time-domain analysis**
 model validation, time-domain approach. *Poolla, K.*, + , *T-AC May 94* 951-959
- Time-optimal control**
 robustness of proximate time-optimal controller. *Pao, L.Y.*, + , *T-AC Sep 94* 1963-1966
- Time series; cf. Autoregressive moving-average processes**
- Time-varying stochastic processes**
 random param. tracking, robust algm. *Juditsky, A.*, + , *T-AC Jun 94* 1211-1221
- Time-varying systems**
 bounded error ident. *Bitanti, S.*, + , *T-AC May 94* 1106-1110
 bounded-error tracking of time-varying params. *Piet-Lahanier, H.*, + , *T-AC Aug 94* 1661-1664
 delay-independent exponential stabil. criteria for time-varying discrete delay systs. *Wu, J.W.*, + , *T-AC Apr 94* 811-814
 discrete-time adaptive controller, global stabil. *Radenkovic, M.S.*, + , *T-AC Nov 94* 2357-2361
 discrete time-varying systs., robust adaptive controller. *Changyun Wen*, *T-AC May 94* 987-991
 exponential lin. quadratic optimal control, discounting. *Hopkins, W.E.*, *Jr.*, *T-AC Jan 94* 175-178
 functional diff. eqns., delay independent stabil. conditions, decay estim. *Lehman, B.*, + , *T-AC Aug 94* 1673-1676
 impulse differential lin. systs., constrained controllability. *Benzaid, Z.*, + , *T-AC May 94* 1064-1066
 infinite time-varying LQ-problem, approx. soln. *Engwerda, J.C.*, *T-AC Jan 94* 235-238
 MIMO stochastic discrete syst. state estim. *Liu Danyang*, + , *T-AC Oct 94* 2087-2091
 nonlin. systs., vibr. control by AP-forcing. *Balestrino, A.*, + , *T-AC Jun 94* 1255-1258
 P-type learning control. *Saab, S.S.*, *T-AC Nov 94* 2298-2302

- quadratic stabilization of continuous time systs. *Mahmoud, M.S.*, +, *T-AC Oct 94* 2135-2139
- random param. tracking, robust algm *Juditsky, A.*, +, *T-AC Jun 94* 1211-1221
- robust stabil., time-varying struct. uncertainty. *Shamma, J.S.*, *T-AC Apr 94* 714-724
- slowly varying and LTI systs., ident and uncertainty principles *Zames, G.*, +, *T-AC Sep 94* 1827-1838
- stabilization, adaptive, nonlin. time-varying controller *Miller, D.E.*, *T-AC Jul 94* 1347-1359
- uncertain nonlin interconnected systs, time-varying state delay, stabilizing control *Mahmoud, M.S.*, *T-AC Dec 94* 2484-2488
- Time-varying systems;** cf. Jump parameter systems, Switched systems
- Toeplitz matrices**
- generalized Chandrasekhar recursions from generalized Schur algm *Sayed, A.H.*, +, *T-AC Nov 94* 2265-2269
- state-space models fast ident *Young Man Cho.*, +, *T-AC Oct 94* 2004-2017
- Topology**
- output regulation, necessary condition *Lucibello, P.*, *T-AC Mar 94* 558-559
- Tracking**
- constrained robots, force/motion control *Grabbe, M.T.*, +, *T-AC Jan 94* 179
- deadbeat ripple-free tracking *Jetto, L.*, *T-AC Aug 94* 1759-1764
- discrete-time nonlin syst control, stabil property *Jie Huang.*, +, *T-AC Nov 94* 2307-2311
- feedback lin. systs adaptive tracking *Marino, R.*, +, *T-AC Jun 94* 1314-1319
- flexible joint robots with uncertain params and disturbances, tracking control *Tomei, P.*, *T-AC May 94* 1067-1072
- MIMO nonlin syst I/O pseudolinarization *Lawrence, D.A.*, +, *T-AC Nov 94* 2207-2218
- MRAC perform anal./improvement, new tracking error criteria *Dattam, A.*, +, *T-AC Dec 94* 2370-2387
- nonlin systs, asymptotic tracking, necessary conditions *Grizzle, J.W.*, +, *T-AC Sep 94* 1782-1794
- nonlin uncertain systs tracking control *Song, Y.D.*, +, *T-AC Sep 94* 1866-1871
- output regulation, necessary condition *Lucibello, P.*, *T-AC Mar 94* 558-559
- prod control methods, distributed, stabil and perform *Sharifnia A.*, *T-AC Apr 94* 725-737
- rigid robotic manipulators, robust tracking control *Man Zhihong.*, +, *T-AC Jan 94* 154-159
- ripple free sampled-data robust servomechanism controller, exponential hold *Yung-Chun Wu.*, +, *T-AC Jun 94* 1287-1291
- robot adaptive control based on passivity *Yu Tang.*, +, *T-AC Sep 94* 1871-1875
- robust stabilization/perform, MRAC modeling error compensation *Sun, J.*, +, *T-AC Mar 94* 630-635
- sampled data control systs and tracking, fn space approach *Yamamoto, Y.*, *T-AC Apr 94* 703-713
- single-arm dyn, robust variable struct./and switching- Σ adaptive control *Li-Wen Chen.*, +, *T-AC Aug 94* 1621-1626
- time-varying params, bounded error ident *Bittanti, S.*, +, *T-AC May 94* 1106-1110
- Tracking;** cf. Radar tracking, Space vehicle tracking
- Tracking loops**
- range tracking loops, large deviation anal *Dembo, A.*, +, *T-AC Feb 94* 360-364
- Transducers;** cf. Multisensor systems
- Transfer function matrices**
- dyn lin systs, pos real lemma generalization *Scherer, R.*, +, *T-AC Apr 94* 882-886
- H_∞ -norm approx of systs by const matrices *Kavranoglu, D.*, *T-AC May 94* 1006-1009
- H_∞ optim, time-domain constraints *Rotstein, H.*, +, *T-AC Apr 94* 762-779
- LTI syst pos real control *Weiqian Sun.*, +, *T-AC Oct 94* 2034-2046
- multivariable lin systs, hysteresis switching MRAC *Weller, S.R.*, +, *T-AC Jul 94* 1360-1375
- neutral systs, stabil and stabilizability *Logemann, H.*, +, *T-AC Jan 94* 138-143
- transfer matrices preconditioning. *Rotstein, B.*, *T-AC Nov 94* 2287-2292
- Transfer functions**
- block multirate input-output model for sampled-data control systs *Jakubowski, A.M.*, +, *T-AC May 94* 1145-1147
- closed loop syst freq response plots, comment. *Gessing, R.*, *T-AC Aug 94* 1770-1771
- closed-loop vibr control, Youla parameterization *Kabamba, P.T.*, +, *T-AC Jul 94* 1455-1459
- continuous coprime factors. *Treil, S.*, *T-AC Jun 94* 1262-1263
- continuous time systs., transfer fns., model errors. *Schoukens, J.*, +, *T-AC Aug 94* 1733-1737
- decentralized stable factors and parameterization of decentralized controllers. *Date, R.*, +, *T-AC Feb 94* 347-351
- delay systs. approx, Laguerre formula. *Lam, J.*, *T-AC Jul 94* 1517-1521
- delta-operator formulated discrete-time approx. *Premaratne, K.*, +, *T-AC Mar 94* 581-585
- discrete-time filters from high-order s-to-z mappings *Schneider, A.M.*, +, *T-AC Feb 94* 435-441
- H_∞ model reduction computational scheme *Kavranoglu, D.*, *T-AC Jul 94* 1447-1451
- H_2 -optimal zeros placement *Kabamba, P.T.*, +, *T-AC Jun 94* 1298-1301
- H_∞ control of discrete-time uncertain systs *Geromel, J.C.*, +, *T-AC May 94* 1072-1075
- lin. syst approx, time-scaling factor, Laguerre models *Wang, L.*, +, *T-AC Jul 94* 1463-1467
- MIMO lin systs with sens./actuator failures, stabil *Gundes, A.N.*, *T-AC Jun 94* 1224-1230
- mixed H_2/H_∞ perform objectives *Kemin Zhou.*, +, *T-AC Aug 94* 1564-1574
- param ident in freq domain *Pintelon, R.*, +, *T-AC Nov 94* 2245-2260
- rationalx transfer fn, optimal L_∞ approx *Kavranoglu, D.*, +, *T-AC Sep 94* 1899-1904
- robust stabilization/perform, MRAC modeling error compensation. *Sun, J.*, +, *T-AC Mar 94* 630-635
- sampled data control systs and tracking, fn. space approach *Yamamoto, Y.*, *T-AC Apr 94* 703-713
- semi-cancellable fraction transfer fns in syst theory *Bourles, H.*, *T-AC Oct 94* 2148-2153
- SISO-distributed plants, optimal mixed sensitivity *Flamm, D.S.*, +, *T-AC Jun 94* 1150-1165
- SISO systs, robust stabilization *Olbro, A.W.*, +, *T-AC Mar 94* 652-657
- SISO systs with affine param uncertainties, robust perform *Kogan, J.*, *T-AC Jan 94* 227-229
- stabilisation, uncertainties/numerator-denominator coupling *Chockalingam, G.*, +, *T-AC Sep 94* 1955-1958
- state delayed systs, memoryless H^∞ controllers *Joon Hwa Lee.*, +, *T-AC Jan 94* 159-162
- state observer/feedback, compensators, comment *Bender, D.J.*, *T-AC Feb 94* 447-448
- state-space models fast ident *Young Man Cho.*, +, *T-AC Oct 94* 2004-2017
- uncertain syst pole assignment, polynomial approach *Figueroa, J.L.*, +, *T-AC Apr 94* 831-835
- uncertain transfer fns, value sets comp *Gutman, P.-O.*, +, *T-AC Jun 94* 1268-1273
- Transforms;** cf. Fourier transforms, Laplace transforms, Wavelet transforms, Z transforms
- Trees, graphs**
- multiscale systs, Kalman filters, Riccati eqns *Chou, K.C.*, +, *T-AC Mar 94* 479-492
- robust anal, uncertainty value sets *Eszter, E.G.*, +, *T-AC Nov 94* 2315-2318
- Tuning**
- Gaussian stochastic control syst tuning *van Schuppen, J.H.*, *T-AC Nov 94* 2178-2190
- Smith predictor for controlling proc., integrator and long dead-time *Astrom, K.J.*, +, *T-AC Feb 94* 343-345
- Two-dimensional systems;** cf. Multidimensional systems

U

Uncertain systems

- cascaded nonlin systs, global robust stabilization. *Imura, J.-I.*, +, *T-AC May 94* 1084-1089
- comments on "Stability margin evaluation for uncertain linear systems" (by C. Gong and S. Thompson, Jun 94 548-550) *Su, J.-H.*, *T-AC Dec 94* 2523-2524
- continuous-time adaptive decoupling control design *Ortega, R.*, +, *T-AC Aug 94* 1639-1643
- design for plants, unknown dead-zones. *Gang Tao.*, +, *T-AC Jan 94* 59-68
- discrete-time lin. systs., pos. invariant sets. *De Santis, E.*, *T-AC Jan 94* 245-249
- discrete-time uncertain systs, ultimate boundedness control, Lyapunov *Blanchini, F.*, *T-AC Feb 94* 428-433
- D-stabil., robust, for lin uncertain discrete delay systs *Te-Jen Su.*, +, *T-AC Feb 94* 425-428
- failure detect./isolation/accommodation syst *Chia-Chi Tsui.*, *T-AC Nov 94* 2318-2321
- H_∞ control of discrete-time uncertain systs *Geromel, J.C.*, +, *T-AC May 94* 1072-1075
- H_∞ control of systs. under norm bounded uncertainties. *Gu, K.*, *T-AC Jun 94* 1320-1322

- interval lin. control syst. robust parametric design. *Keel, L.H.*, +, *T-AC Jul 94* 1524-1530
- lin. quadratic designs, real param. uncertainty, robust. *Douglas, J.*, +, *T-AC Jan 94* 107-111
- MRAC, robust, SISO systs. *Zhihua Qu.*, +, *T-AC Nov 94* 2219-2234
- nonlin. interconnected systs., time-varying state delay, stabilizing control. *Mahmoud, M.S.*, *T-AC Dec 94* 2484-2488
- nonlin. uncertain systs. tracking control. *Song, Y.D.*, +, *T-AC Sep 94* 1866-1871
- optimal cost control/filtering, uncertain lin. systs. *Petersen, I.R.*, +, *T-AC Sep 94* 1971-1977
- optimal filtering, stochastic discrete-time systs., unknown inputs. *Bortsov, A.V.*, +, *T-AC Dec 94* 2461-2464
- parametrically uncertain nonlin. feedback systs., absol. stabil. *Marquez, H.J.*, +, *T-AC Mar 94* 664-668
- parametrically uncertain systs., anal. and synthesis. *Kaminsky, R.D.*, +, *T-AC Apr 94* 874-876
- quadratic stabilization of continuous time systs. *Mahmoud, M.S.*, +, *T-AC Oct 94* 2135-2139
- robust anal., uncertainty value sets. *Eszter, E.G.*, +, *T-AC Nov 94* 2315-2318
- robust stabil., time-varying struct. uncertainty. *Shamma, J.S.*, *T-AC Apr 94* 714-724
- SISO systs. with affine param. uncertainties, robust perform. *Kogan, J.*, *T-AC Jan 94* 227-229
- slowly varying and LTI systs., ident. and uncertainty principles. *Zames, G.*, +, *T-AC Sep 94* 1827-1838
- stabilisation, uncertainties/numerator-denominator coupling. *Chockalingam, G.*, +, *T-AC Sep 94* 1955-1958
- stabil. of uncertain lin. systs., saturating actuators. *Jin-Hoon Kim.*, +, *T-AC Jan 94* 202-207
- structurally uncertain systs., optimal Lyapunov fns. *Olas, A.*, *T-AC Jan 94* 167-171
- time-invariant lin. systs. controllability/observability. *Kaining Wang.*, +, *T-AC Jul 94* 1443-1447
- uncertain continuous-time implicit systs., regional pole placement, robustness. *Chun-Hsiung Fang.*, +, *T-AC Nov 94* 2303-2307
- uncertain discrete-time systs., variable struct. control design. *Myszkowski, P.*, +, *T-AC Nov 94* 2366-2367
- uncertain lin. systs., delay depend., robust stabil. *Bugong Xu.*, *T-AC Nov 94* 2365
- uncertain lin. systs., multivariable, stabil. margin eval. *Gong, C.*, +, *T-AC Mar 94* 548-550
- uncertain nonlin. systs., struct. invariance. *Castro-Linares, R.*, +, *T-AC Oct 94* 2154-2158
- uncertain SISO min. phase lin. syst., nonlin. universal servomechanism. *Ryan, E.P.*, *T-AC Apr 94* 753-761
- uncertain syst. pole assignment, polynomial approach. *Figuerola, J.L.*, +, *T-AC Apr 94* 831-835
- uncertain syst. robust stabil. anal., nonquadratic Lyapunov fns. *Zelentsovsky, A.L.*, *T-AC Jan 94* 135-138
- uncertain syst. robust stabil., guardian map approach. *Shuoh Rern.*, +, *T-AC Jan 94* 162-164
- uncertain systs., asymptotic stabil. region estim. *Han Ho Chol.*, +, *T-AC Nov 94* 2275-2278
- uncertain systs., sampled-data controller design. *Dolphus, R.M.*, *T-AC May 94* 1036-1042
- uncertain systs. stabilizing control design, quasiconvex optim. *Kegin Gu.*, *T-AC Jan 94* 127-131
- unknown transfer fns., value sets comp. *Gutman, P.-O.*, +, *T-AC Jun 94* 1268-1273
- unknown input observers design. *Darouach, M.*, *T-AC Mar 94* 698-699
- VSS control design for uncertain discrete-time systs. *Wen-June Wang.*, +, *T-AC Jan 94* 99-102
- μ calc., computational complexity. *Braatz, R.P.*, +, *T-AC May 94* 1000-1002

V

Variable-structure systems

- constrained robots, force/motion control. *Grabbe, M.T.*, +, *T-AC Jan 94* 179

- discrete-time uncertain systs., ultimate boundedness control, Lyapunov. *Blanchini, F.*, *T-AC Feb 94* 428-433
- generator voltage regulator, pole assignment tech. *Aggoune, M.E.*, +, *T-AC Oct 94* 2106-2110
- MRAC, variable struct. I/O based, anal./design. *Liu Hsu.*, +, *T-AC Jan 94* 4-21
- rigid robotic manipulators, robust tracking control. *Man Zhihong.*, +, *T-AC Jan 94* 154-159
- robot variable struct. control schemes. *Bin Yao.*, +, *T-AC Feb 94* 371-376
- robust control design, cascade struct. approach. *Bonivento, C.*, +, *T-AC Apr 94* 846-849
- robust MIMO terminal sliding-mode control, rigid robotic manipulators. *Zhihong, M.*, +, *T-AC Dec 94* 2464-2469
- single-arm dyn., robust variable struct./and switching- Σ adaptive control. *Li-Wen Chen.*, +, *T-AC Aug 94* 1621-1626
- switching surface design for multivariable VSS. *Ju-Jang Lee.*, +, *T-AC Feb 94* 414-419
- uncertain discrete-time systs., variable struct. control design. *Myszkowski, P.*, +, *T-AC Nov 94* 2366-2367
- uncertain discrete-time systs., VSS control design. *Wen-June Wang.*, +, *T-AC Jan 94* 99-102
- uncertain systs., asymptotic stabil. region estim. *Han Ho Chol.*, +, *T-AC Nov 94* 2275-2278

Vectors

- comments, with reply, on "Vector norms as Lyapunov functions for linear systems" (by H. Kiendl et al., Jun 92 839-842). *Hmamed, A.*, *T-AC Dec 94* 2522-2523

Velocity control

- robust control design, cascade struct. approach. *Bonivento, C.*, +, *T-AC Apr 94* 846-849

Very-large-scale integration

- constrained nonlin. syst. receding horizon control. *Shin-Yeu Lin.*, *T-AC Sep 94* 1893-1899

Vibration control

- closed-loop vibr. control, Youla parameterization. *Kabamba, P.T.*, +, *T-AC Jul 94* 1455-1459
- nonlin. uncertain systs. tracking control. *Song, Y.D.*, +, *T-AC Sep 94* 1866-1871
- second-order dyn. systs., dissipative controller designs. *Morris, K.A.*, +, *T-AC May 94* 1056-1063
- structural wave control, reduced-order model. *Quan Wang.*, +, *T-AC Aug 94* 1711-1713
- time lag nonlin. systs., vibr. control. *Lehman, B.*, +, *T-AC May 94* 898-912
- Vision systems (non-biological); cf. Robots, vision systems
- VLSI; cf. Very-large-scale integration

W

Wavelet transforms

- multiresolutional distributed filtering. *Lang Hong.*, *T-AC Apr 94* 853-856

White noise

- least sq. estim. in white noise, convergence. *Nassiri-Toussi, K.*, +, *T-AC Feb 94* 364-368
- model struct. selection test, instrumental variable, statist. props. *Hoai Nghia Duong.*, +, *T-AC Jan 94* 211-215

Wiener processes

- recursive ident., nonlin. Wiener model. *Wigren, T.*, *T-AC Nov 94* 2191-2206
- Wiener systs. nonparametric ident. *Greblicki, W.*, *T-AC Oct 94* 2077-2086

Z

Zero assignment

- H₂-optimal zeros placement. *Kabamba, P.T.*, +, *T-AC Jun 94* 1298-1301

Z transforms

- transfer fn. param. ident. in freq. domain. *Pintelon, R.*, +, *T-AC Nov 94* 2245-2260

Scanning the Issue*

Adjoint and Hamiltonian Input-Output Differential Equations, *Crouch, Lamnabhi-Lagarigue, and van der Schaft*.

In 1887, Helmholtz established conditions under which a set of second-order nonlinear differential equations are, or can be transformed into, Lagrangian (or Hamiltonian) form. This paper generalizes the classical Helmholtz conditions to nonlinear input-output differential systems, i.e., to nonlinear control systems. The adjoint variational system and the meaning of self-adjointness for a general input-output system are defined in the course of the analysis. The results employ recent developments in the theory of variational and Hamiltonian control systems. A main advantage of the results of this paper over previous related results is their verifiability directly in terms of the given input-output model. Several examples are given illustrating the results and associated computations.

Language Convergence in Controlled Discrete-Event Systems, *Willner and Heymann*.

This paper addresses the issue of how to supervise a discrete-event system modeled by a finite automaton so that after a bounded, but unknown, number of transitions, the behavior of the controlled system will conform to some specified behavior. Several issues within this problem area are addressed. Necessary and sufficient conditions for finite convergence are provided, and algorithms for implementing these tests are developed. These are then applied to achieving language convergence through control. This is discussed within the context of finite string behavior as well as asymptotic behavior. The main contribution is a detailed analysis of existence and computation of supervisors to achieve eventually correct supervisory control.

Concurrent Vector Discrete-Event Systems, *Li and Wonham*.

In earlier papers published in these transactions, the authors have addressed the control of vector discrete-event systems (VDES). VDES are a class of discrete-event models where the state is represented by a vector of integers and state transitions are manifested through integer vector addition. This paper extends VDES control theory by incorporating simultaneity of events into the model and by considering nondeterminism of controllers. The extended VDES model with simultaneous (or concurrent) events is termed concurrent VDES (CVDES). The main part of the paper deals with nondeterministic control of CVDES. The nondeterminism comes from the fact that the control action issued by the controller at each step is "randomly" (i.e., according to some hidden mechanism) chosen among a set of possible control actions. The authors show how to exploit the power of nondeterministic control in the presence of concurrency.

Structure of Model Uncertainty for a Weakly Corrupted Plant, *Zhou and Kimura*.

Robust control paradigms are typically based on a plant model set described by a nominal model and a bound on the distance, in a certain metric, between this nominal model and all other models in the set. For paradigms to be usable, a procedure must be available to determine an appropriate model set from available plant information. Several philosophies have been put forth that assign different meanings to "appropriateness." Consistent with the control motivation, the philosophy embraced in the present paper is that of

"worst case deterministic" modeling, by which the model set is to be the smallest set of plants which includes all those which could have produced the available plant information, under given assumptions on the noise, and which satisfy given *a priori* conditions. Specifically, the first task undertaken by the authors is to characterize the set of all linear, time-invariant, stable transfer function matrices, with the H_∞ norm below a prescribed threshold, which are consistent with given experimental data known to be weakly affected by noise. The key result here is a parameterization of this set as the linear fractional transformation of a fixed transfer function matrix and a structured, norm bounded, free transfer function matrix. Next, with the robust control problem in mind, the authors consider determining the smallest superset of the set just characterized that consists of a "nominal" transfer function matrix of prescribed order, and of an H_∞ "ball" around it. They show that this problem can be expressed as a mixed- m synthesis problem.

Explicit Formulas for Optimally Robust Controllers for Delay Systems, *Dym, Georgiou, and Smith*.

H_∞ control for distributed parameter systems, initiated in the mid 1980's, focused first on developing formulas to provide further insight into limitations on performance due to infinite dimensional elements (such as time-delays) and second, on developing a direct approach to design which does not rely on first finding a finite dimensional approximation of the plant. Such an approach has the potential to shed light on the structure of ideal compensators for distributed parameter systems.

In this paper, the authors consider single-input/single-output systems whose transfer functions consist of a strictly proper rational function times a delay. They present a closed-form expression for the controller which is optimally robust with respect to perturbations of the nominal plant measured in the gap metric. The derivations involve a synthesis of state-space techniques and the use of a certain algebra of "pseudo-derivation" operators, which amounts to a specific class of distributed delays. The controller transfer function has a simple expression involving state-space matrices computed from the rational part of the plant together with a distributed delay element. The expression makes it quite straightforward to carry out directly the H_∞ loop-shaping procedure of Glover-McFarlane for the class of plants considered and rational weights. An example design is worked out to illustrate the technique and a comparison is made with a Smith predictor.

Stochastic System Identification with Noisy Input-Output Measurements Using Polyspectra, *Tugnait and Ye*.

In this paper, parameter estimation in linear SISO systems (with prescribed orders) is considered. Here inputs and outputs are assumed to be contaminated by noise, and identifiability is obtained by additional assumptions on the probability distributions of the signal and noise processes, respectively. Based on a simple relation between the true transfer function and some (population) bispectra or integrated polyspectra, first estimators of the transfer function are derived from the corresponding sample counterpoints. From these, two classes of estimators for the transfer function, one called linear (which, strictly speaking is also nonlinear) and one called nonlinear, are derived. Consistency is analyzed, and bias and variance of the estimators are compared (also with other, already existing, estimators) by means of a Monte-Carlo simulation.

* This section is written by the Transactions Editorial Board.

Minimum Bias Priors for Estimating Parameters of Additive Terms in State-Space Models, *Hochwald and Nehorai*.

We treat the problem of estimating parameters of additive terms, sometimes called bias terms, in state-space models. We consider

models that depend linearly on the state but possibly nonlinearly on the parameters, where both the state and observation are corrupted by additive noise. A prior density for the parameters is introduced that, when combined with the likelihood function to form a posterior density, minimizes the bias of the posterior mean.

Adjoint and Hamiltonian Input–Output Differential Equations

Peter E. Crouch, *Senior Member, IEEE*, Francoise Lamnabhi-Lagarrigue, and Arjan J. van der Schaft

Abstract—Based on recent developments in the theory of variational and Hamiltonian control systems by Crouch and van der Schaft, this paper answers two questions: given an input–output differential equation description of a nonlinear system, what is the adjoint variational system in input–output differential form and what are the conditions for the system to be Hamiltonian, i.e., such that the variational and the adjoint variational systems coincide? This resulting set of conditions is then used to generalize classical conditions such as the well-known Helmholtz conditions for the inverse problem in classical mechanics.

I. INTRODUCTION

THE work we are describing in this paper has its roots in a very old problem in classical mechanics, where one asks which Newtonian systems correspond to Lagrangian, or variational systems; the so-called inverse problem. There are many variants of this problem, see Santilli [14], but the simplest one can be stated as follows.

If $q \in \mathbb{R}^n$ is a configuration variable, which satisfies the Newtonian system

$$F(q, \dot{q}, \ddot{q}) = 0 \quad (1)$$

for a smooth \mathbb{R}^n valued mapping F , such that $\partial F / \partial \ddot{q}$ is a nonsingular matrix in a suitable open domain, then under what conditions does there exist a function L of q, \dot{q} , so that for some ordering of the variables of q

$$F(q, \dot{q}, \ddot{q}) = \frac{d}{dt} \frac{\partial L^T}{\partial \dot{q}}(q, \dot{q}) - \frac{\partial L^T}{\partial q}(q, \dot{q}). \quad (2)$$

The conditions under which this property holds are known as the classical Helmholtz conditions, see Santilli [14] where generalizations to functions F depending on arbitrary finite jets of q are also considered.

The condition on the rank of $\partial F / \partial \ddot{q}$ in the system (1) above enables those systems satisfying (2) to be written also as a

Hamiltonian set of equations

$$\begin{aligned} \dot{q} &= \frac{\partial H^T}{\partial p}(q, p) \\ \dot{p} &= -\frac{\partial H^T}{\partial q}(q, p) \end{aligned} \quad (3)$$

where $H(p, q)$ is the Hamiltonian function of the system, and (p, q) are coordinates on the symplectic phase space \mathbb{R}^{2n} . A result of Brockett and Rahimi [1] is also of interest in this context and concerns the linear system Σ

$$\begin{aligned} \dot{x} &= Ax + Bu; \quad x(0) = 0 \\ y &= Cx \end{aligned} \quad (4)$$

in which $x \in \mathbb{R}^n$, $u, y \in \mathbb{R}^m$, and the so-called adjoint system Σ^a (see also [12])

$$\begin{aligned} \dot{p} &= -A^T p - C^T u_a; \quad p(0) = 0 \\ y_a &= B^T p. \end{aligned} \quad (5)$$

It was shown with the minimality of both systems, together with the “self-adjointness” condition that the input–output maps of Σ and Σ^a coincide, that this is equivalent to the fact that the system Σ has another internal representation as a linear “Hamiltonian Control” system

$$\begin{aligned} \dot{p} &= -\frac{\partial H^T}{\partial q}(p, q, u) \\ \dot{q} &= \frac{\partial H^T}{\partial p}(p, q, u) \\ y &= \frac{\partial H^T}{\partial u}(p, q, u) \end{aligned} \quad (6)$$

where for a linear Hamiltonian system

$$H(p, q, u) = \frac{1}{2}(p^T, q^T)F \begin{bmatrix} p \\ q \end{bmatrix} + (p^T, q^T)Gu$$

for some matrices F and G .

The term “self-adjointness” is also used to describe the conditions which ensure that a Newtonian system does correspond to a Lagrangian, or Hamiltonian, system. This is explained by performing integration by parts to give an expression

$$\begin{aligned} &\int_0^T \xi^T \left[\frac{\partial F}{\partial q} q_v + \frac{\partial F}{\partial \dot{q}} \dot{q}_v + \frac{\partial F}{\partial \ddot{q}} \ddot{q}_v \right] dt \\ &= \int_0^T q_v^T F^*(\xi, \dot{\xi}, \ddot{\xi}) dt \\ &\quad + Q(q, \dot{q}, \ddot{q}, q_v, \dot{q}_v, \ddot{q}_v, \xi, \dot{\xi}, \ddot{\xi}) \Big|_0^T \end{aligned} \quad (7)$$

Manuscript received November 26, 1991; revised January 3, 1994 and August 9, 1994. Paper recommended by Past Associate Editor, E. H. Abed. The work of P. E. Crouch was supported in part by the National Science Foundation under Grants INT 8914643 and DMS 9101964. The work of F. Lamnabhi-Lagarrigue was supported in part by the National Science Foundation under Grant 0693 and by EEC Science RTD SCI0433CA.

P. E. Crouch is with the Center for Systems Science and Engineering, Arizona State University, Tempe, AZ 85287 USA.

F. Lamnabhi-Lagarrigue is with Laboratoire des Signaux et Systèmes, SUPELEC, 91192 Gif-sur-Yvette Cedex, France.

A. J. Van der Schaft is with the Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

IEEE Log Number 9408274.

for some functions F^* and Q . The adjoint variational system to the variational system $\ddot{q}_v + \frac{\partial F}{\partial q} \dot{q}_v + \frac{\partial F}{\partial \dot{q}} \ddot{q}_v = 0$ corresponding to (1) is then

$$F^*(\xi, \dot{\xi}, \ddot{\xi}) = 0.$$

Self-adjointness of F is simply the statement that

$$F^*(\xi, \dot{\xi}, \ddot{\xi}) = \frac{\partial F}{\partial q} \xi + \frac{\partial F}{\partial \dot{q}} \dot{\xi} + \frac{\partial F}{\partial \ddot{q}} \ddot{\xi}.$$

Work by Crouch and van der Schaft [5], [6], made an extensive investigation and generalization of the result by Brockett and Rahimi [1] to nonlinear control systems, based on earlier work by van der Schaft [15]. In particular, a state space form of the variational and adjoint variational systems was introduced, generalizing the relationship between Σ and Σ^a to nonlinear systems. The concept of self-adjointness was correspondingly generalized, and under suitable hypotheses it was shown that self-adjointness is necessary and sufficient for Hamiltonian realizations of input-output maps. Moreover, the self-adjointness condition was successfully interpreted in terms of the Volterra series and Fliess series. See also Jakubczyk [12] for a generalization to control systems with control entering in a nonaffine manner.

This work, although implicitly generalizing the classical Helmholtz condition, fails to give conditions in terms of the differential equation representation of the input-output map, generalizing the system representation given by (1), and here represented by an equation of the form

$$F(y, y^{(1)}, \dots, y^{(N)}, u, u^{(1)}, \dots, u^{(N-1)}) \equiv 0 \quad (8)$$

where F is a smooth vector valued function of its arguments. Furthermore, the self-adjointness conditions of [5] and [6] are difficult to check in practice, since in principle one needs to compute the state trajectories of the nonlinear system under consideration. The present paper gives the full generalization of the classical Helmholtz conditions to control systems described by (8). The resulting conditions are completely in terms of the mapping F and its partial derivatives. They are worked out in detail for control systems

$$F(y, y^{(1)}, y^{(2)}, u, u^{(1)}) = 0 \quad (9)$$

so that they have a Hamiltonian representation by a system of the form (6). To the knowledge of the authors, the only previous results in this direction are those given by two of the current authors [4], when dealing with the scalar input-scalar output version of the system (9). It is interesting to note that many problems associated with the preceding analysis coincide with those met in the study of time-varying linear systems; see, e.g., [2], [9], [10], [11], and [13].

II. THE ADJOINT VARIATIONAL SYSTEM

We consider analytic (i.e., C^ω), complete, state-space systems which may be written in the form

$$\Sigma_s \quad \dot{x} = f(x, u), \quad y = h(x), \quad u \in \mathbb{R}^m, \quad y \in \mathbb{R}^p \quad (10)$$

where $x = (x_1, \dots, x_n)$ denote local coordinates for some state space manifold M and corresponding analytic input-output differential representations

$$\Sigma_{i/o} \quad F(y, \dot{y}, \dots, y^{(N)}, u, \dot{u}, \dots, u^{(N-1)}) = 0 \in \mathbb{R}^p. \quad (11)$$

If system (10) is minimal in the sense of Crouch and van der Schaft [5, ch. 3 for the input-affine case and ch. 6 for the general case], the corresponding representation (11) will also be called a minimal representation. Note that we do not insist here on the relationship between state-space representations (10) and input-output differential representations (11); we assume that all conditions are met for obtaining one representation from the other representation.

A variational system Σ_s^v about a given trajectory $(x(t), u(t), y(t))$ of Σ_s is defined in the usual way (see [5]) as

$$\begin{aligned} \dot{v}(t) &= F(t)v(t) + G(t)u_v(t), & u_v &\in \mathbb{R}^m, \quad v \in \mathbb{R}^n \\ y_v(t) &= H(t)v(t), & y_v &\in \mathbb{R}^p \end{aligned} \quad (12)$$

with $F(t) := \frac{\partial f}{\partial x}(x(t), u(t))$, $G(t) := \frac{\partial f}{\partial u}(x(t), u(t))$, $H(t) = \frac{\partial h}{\partial x}(x(t))$, and v , u_v , y_v denoting, respectively, the variational state, variational input, and variational output. Note that (12) results from differentiation of a one-parameter family of solutions to (10), cf. [5]. The adjoint variational system Σ_s^a , along the same trajectory $(x(t), u(t), y(t))$, is defined as (see [5])

$$\begin{aligned} \dot{p}(t) &= -F^T(t)p(t) - H^T(t)u_a(t), & u_a &\in \mathbb{R}^p, \quad p \in \mathbb{R}^n \\ y_a(t) &= G^T(t)p(t), & y_a &\in \mathbb{R}^m \end{aligned} \quad (13)$$

with p , u_a , y_a denoting the adjoint variational state, input, and output, respectively. The fundamental connection between variational and adjoint variational systems (along the same trajectory of Σ_s) is [5, Lemma 2.1]

$$y_a^T(t)u_v(t) - u_a^T(t)y_v(t) = \frac{d}{dt}p^T(t)v(t). \quad (14)$$

Now let us translate this to input-output differential representations $\Sigma_{i/o}$ given by (11). Clearly, the variational systems $\Sigma_{i/o}^v$ (along solutions $u(t), y(t)$ of $\Sigma_{i/o}$) are defined by the system of equations

$$\begin{aligned} \frac{\partial F}{\partial y} y_v(t) + \frac{\partial F}{\partial \dot{y}} \dot{y}_v(t) + \dots + \frac{\partial F}{\partial y^{(N)}} y_v^{(N)}(t) + \frac{\partial F}{\partial u} u_v(t) \\ + \frac{\partial F}{\partial \dot{u}} \dot{u}_v(t) + \dots + \frac{\partial F}{\partial u^{(N-1)}} u_v^{(N-1)}(t) = 0 \end{aligned} \quad (15)$$

where the solution $u(t), y(t)$ is substituted in $\frac{\partial F}{\partial y}$, $\frac{\partial F}{\partial \dot{y}}$, \dots , $\frac{\partial F}{\partial u}$, $\frac{\partial F}{\partial \dot{u}}$, \dots . Again, (15) results from differentiation of a one-parameter family of solutions $(u(t, c), y(t, c))$ to (11). Comparing $\Sigma_{i/o}^v$ to Σ_s^v there is a potential problem since (see [20]) the set of solutions u_v, y_v to (15) may be strictly larger than the set of solutions u_v, y_v generated by (12). This has to do with the form of the input-output differential representation (11) (note that this representation is far from unique); we will later on make an assumption on (11) which will eliminate this potential problem.

The next logical question is how to define the adjoint variational system $\Sigma_{i/o}^a$. This is not immediate from the

definition of Σ_s^a . Relation (14) should provide the clue to a proper definition, although the right-hand side of (14) is expressed in the variational and adjoint variational state. Direct calculation based on Σ_s^v and Σ_s^a provides the following alternative to (14)

$$\begin{aligned} y_a^T(t)u_v(t) - u_a^T(t)y_v(t) \\ = \frac{d}{dt} \left[- \int_{-\infty}^t \int_{-\infty}^s u_a^T(\tau) H(\tau) \Phi(\tau, s) G(s) u_v(s) ds d\tau \right] \end{aligned} \quad (16)$$

with $\Phi(\tau, s)$ being the transition matrix of $F(t)$ (and the variational and adjoint variational states initialized at 0 at time $-\infty$). Although the right-hand side of (16) is an integral expression in u_a, u_v , this partly motivates the following definition (another motivation is provided by the classical definition of adjoint variational systems for sets of differential equations, cf. [14]).

Definition 1 (see also [5]): Consider the variational system (15) along a solution $(u(t), y(t))$ of $\Sigma_{i/o}$. The adjoint variational system $\Sigma_{i/o}^a$ consists of the following set of input-output pairs $(u_a(\cdot), y_a(\cdot))$: there exists a function $Q(y, \dot{y}, \dots, u, \dot{u}, \dots, y_v, \dot{y}_v, \dots, u_v, \dot{u}_v, \dots, y_a, \dot{y}_a, \dots, u_a, \dot{u}_a, \dots)$ such that for all $t \in R$

$$y_a^T(t)u_v(t) - u_a^T(t)y_v(t) = \frac{d}{dt} Q \quad (17)$$

for all input-output pairs $(u_v(\cdot), y_v(\cdot))$ which are solutions of $\Sigma_{i/o}^v$.

To justify this definition we first have to prove that Definition 1 uniquely characterizes the adjoint variational system.

Proposition 1: Suppose that there exists a function

$$\begin{aligned} \hat{Q}(y, \dot{y}, \dots, u, \dot{u}, \dots, y_v, \dot{y}_v, \dots, u_v, \dot{u}_v, \dots, \\ \tilde{y}_a, \dot{\tilde{y}}_a, \dots, u_a, \dot{u}_a, \dots) \end{aligned}$$

such that for all solutions $(u_v(t), y_v(t))$ to $\Sigma_{i/o}^v$ and all $u_a(t)$

$$\tilde{y}_a^T(t)u_v(t) - u_a^T(t)y_v(t) = \frac{d}{dt} \hat{Q} \quad (18)$$

for all $t \in R$, then $\tilde{y}_a(t) = y_a(t)$, $t \in R$, and $\hat{Q} = Q$ modulo a constant.

Proof. Subtracting (17) from (18) yields

$$(\tilde{y}_a - y_a)^T u_v = \frac{d}{dt} (\hat{Q} - Q)$$

for all u_v . Hence

$$\int_{t_1}^t [\tilde{y}_a(t) - y_a(t)]^T u_v(t) dt = [\hat{Q} - Q]_{t_1}^t. \quad (19)$$

Now take a fixed function u_a on $[t_1, t_2]$ and corresponding fixed functions y_a, \tilde{y}_a on $[t_1, t_2]$ and arbitrary u_v . It follows that the left-hand side of (19) only depends on $u_v(t_1)$ and $u_v(t_2)$ (since the right-hand side does), thus implying that both sides of (19) are zero, and $\tilde{y}_a = y_a$ and $\hat{Q} = Q$ modulo a constant. ■

The next thing we have to do is to show that Definition 1 is consistent with the definition of the adjoint variational system Σ_s^a . Comparing (17) to (14) we see that this

means that $p^T v$ has to be expressible as a function Q of $y_v, \dot{y}_v, \dots, u_v, \dot{u}_v, \dots, y_a, \dot{y}_a, \dots, u_a, \dot{u}_a, \dots$ (and of course $y, \dot{y}, \dots, u, \dot{u}, \dots$). To do so we make fundamental use of some results obtained by Ilchmann *et al.* [10] on time-varying linear systems and Coron [3] and Sontag [17] on the relation between nonlinear state space systems Σ_s and the variational systems Σ_s^v ; see also [3], [18], and [7]. Indeed, in [3], [17] the following is shown. Consider the minimal state space system Σ_s . Let I be an open interval of R , and denote by $C^\infty(I; R^m)$ the set of smooth input functions $u: I \rightarrow R^m$, equipped with the Whitney topology. Then the set of all u in $C^\infty(I; R^m)$ such that all corresponding solutions $(x(t), u(t))$ of Σ_s defined on I have the property that the variational systems (12) along $(x(t), u(t))$ satisfy

$$\dim \text{span} \left\{ \left(\frac{d}{dt} - F(t) \right)^i G(t) w; \quad w \in R^m, i \geq 0 \right\} = n \quad (20)$$

for all $t \in I$, is a dense subset of $C^\infty(I; R^m)$; see [3, Corollary 1.8]. (Note that the definition of the strong accessibility algebra used in [3] is the usual definition in the case of input-affine systems Σ_s , while for general systems $\dot{x} = f(x, u)$ it corresponds exactly to the definition given in [5, ch. 6].)

Furthermore, see [3, Corollary 1.15], the set of all u in $C^\infty(I; R^m)$ such that all corresponding solutions $(x(t), u(t))$ of Σ_s defined on I have the property that the variational systems (12) along $(x(t), u(t))$ satisfy

$$\dim \text{span} \left\{ \left(\frac{d}{dt} + F^T(t) \right)^i H^T(t) w; \quad w \in R^p, i \geq 0 \right\} = n \quad (21)$$

for all $t \in I$, is also a dense subset of $C^\infty(I; R^m)$. Properties (20) and (21) express well-known controllability, respectively, observability, properties of the time-varying linear systems given by the variational systems Σ_s^v , and thus the above statements imply, loosely speaking, that in case Σ_s is minimal then its variational systems are controllable and observable for a dense subset of input functions.

To use now the fundamental results obtained in Ilchmann *et al.* [10] and Ilchmann [11] we will now restrict to analytic (C^ω) input functions on the time-interval I . Since (10) is assumed to be analytic this will mean that the variational and adjoint variational systems (12), respectively (13), are analytic, i.e., the entries of $F(t)$, $G(t)$, and $H(t)$ are analytic functions on I . Let \mathcal{M} denote the meromorphic functions on I , and denote by $\mathcal{M}[D]$ the set of polynomials

$$\sum_{i=0}^n f_i D^i$$

in D with coefficients from \mathcal{M} (D will represent the differentiation operator $\frac{d}{dt} \in \text{End}(\mathcal{M})$, the algebra of R -linear maps from \mathcal{M} to \mathcal{M}). Considering also the multiplication in $\text{End}(\mathcal{M})$, we arrive at the skew-polynomial ring $\mathcal{M}[D]$ (see [10], [11]) with multiplication rule

$$\begin{aligned} D(fg) &= fD(g) + D(f)g = (fD + D(f))g, \\ f, g &\in \text{End}(\mathcal{M}). \end{aligned} \quad (22)$$

The set of $m \times n$ matrices over $\mathcal{M}[D]$ will be denoted by $\mathcal{M}[D]^{m \times n}$. A useful property (see [10]) is that a left inverse of a square matrix in $\mathcal{M}[D]^{n \times n}$ is also a right inverse, and vice versa.

Let us now denote the variational and adjoint variational system in shorthand notation by

$$\begin{aligned} \left(I \frac{d}{dt} - F\right)v - Gu_v &= 0 \\ y_v - Hv &= 0 \end{aligned} \quad (23)$$

$$\begin{aligned} \left(I \frac{d}{dt} + F^T\right)p + H^T u_a &= 0 \\ y_a - G^T p &= 0. \end{aligned} \quad (24)$$

Throughout we will restrict to the dense subset of analytic input functions for which (23) satisfies the controllability and observability properties (20) and (21). Then by [10, Lemma 3.3 and Theorem 6.4] there exists an invertible matrix

$$\begin{bmatrix} \tilde{U}(\frac{d}{dt}) & \tilde{W}(\frac{d}{dt}) \\ \tilde{V}(\frac{d}{dt}) & \tilde{Z}(\frac{d}{dt}) \end{bmatrix} \in \mathcal{M}[D]^{(n+m) \times (n+m)}$$

such that

$$\left[I \frac{d}{dt} - F, -G\right] \begin{bmatrix} \tilde{U}(\frac{d}{dt}) & \tilde{W}(\frac{d}{dt}) \\ \tilde{V}(\frac{d}{dt}) & \tilde{Z}(\frac{d}{dt}) \end{bmatrix} = [I_n, 0] \quad (25)$$

and an invertible matrix

$$\begin{bmatrix} \tilde{P}(\frac{d}{dt}) & \tilde{Q}(\frac{d}{dt}) \\ \tilde{S}(\frac{d}{dt}) & \tilde{R}(\frac{d}{dt}) \end{bmatrix} \in \mathcal{M}[D]^{(n+p) \times (n+p)}$$

such that

$$\left[I \frac{d}{dt} + F^T, H^T\right] \begin{bmatrix} \tilde{P}(\frac{d}{dt}) & \tilde{Q}(\frac{d}{dt}) \\ \tilde{S}(\frac{d}{dt}) & \tilde{R}(\frac{d}{dt}) \end{bmatrix} = [I_n, 0]. \quad (26)$$

We will now apply the following "integration by parts" procedure to (25), (26). This will be a key tool in all of our subsequent developments.

Consider two matrices $M(\frac{d}{dt})$, $N(\frac{d}{dt})$ over $\mathcal{M}[D]$, of appropriate dimensions such that

$$M\left(\frac{d}{dt}\right)N\left(\frac{d}{dt}\right) = C \quad (27)$$

with C some constant matrix, say of dimension $k \times l$. (Note that $M(\frac{d}{dt})$ acts also on the entries of $N(\frac{d}{dt})$, cf. (22); the same of course applies to the expressions (25) and (26).)

Let now $\xi(t)$ be an l -vector of analytic functions, and consider the k -vector of analytic functions

$$M\left(\frac{d}{dt}\right)N\left(\frac{d}{dt}\right)\xi(t). \quad (28)$$

Now premultiply (28) by a k -row vector of analytic functions $\eta^T(t)$, i.e.,

$$\eta^T(t)M\left(\frac{d}{dt}\right)N\left(\frac{d}{dt}\right)\xi(t) \quad (29)$$

and apply integration by parts to the integral

$$\int_{t_1}^{t_2} \eta^T(t)M\left(\frac{d}{dt}\right)N\left(\frac{d}{dt}\right)\xi(t) dt$$

(with $[t_1, t_2] \subset I$) in order to shift the differentiations on $\xi(t)$ to differentiations on $\eta(t)$. It follows that

$$\begin{aligned} \int_{t_1}^{t_2} \eta^T(t)M\left(\frac{d}{dt}\right)N\left(\frac{d}{dt}\right)\xi(t) dt \\ = \int_{t_1}^{t_2} \xi^T(t)\hat{N}\left(\frac{d}{dt}\right)\tilde{M}\left(\frac{d}{dt}\right)\eta(t) dt + \text{remainders} \end{aligned} \quad (30)$$

for certain matrices $\hat{N}(\frac{d}{dt})$, $\tilde{M}(\frac{d}{dt})$ over $\mathcal{M}[D]$ (in fact, \hat{N} and \tilde{M} are dimensioned as M^T , respectively N^T). Furthermore, from (27) and (30) it follows that the differential operator $\hat{N}(\frac{d}{dt})\tilde{M}(\frac{d}{dt})$ equals the constant linear mapping C^T on all functions $\eta(t)$ of support within (t_1, t_2) . However, this means that $\hat{N}(\frac{d}{dt})\tilde{M}(\frac{d}{dt})$ equals C^T , and thus the remainders in (30) are necessarily zero.

Applying this procedure to (25) and (26) yields

$$\begin{bmatrix} U(\frac{d}{dt}) & V(\frac{d}{dt}) \\ W(\frac{d}{dt}) & Z(\frac{d}{dt}) \end{bmatrix} \begin{bmatrix} -I \frac{d}{dt} - F^T \\ -G^T \end{bmatrix} = \begin{bmatrix} I_n \\ 0 \end{bmatrix} \quad (31)$$

$$\begin{bmatrix} P(\frac{d}{dt}) & S(\frac{d}{dt}) \\ Q(\frac{d}{dt}) & R(\frac{d}{dt}) \end{bmatrix} \begin{bmatrix} -I \frac{d}{dt} + F \\ H \end{bmatrix} = \begin{bmatrix} I_n \\ 0 \end{bmatrix} \quad (32)$$

respectively, for invertible matrices $\begin{bmatrix} U & V \\ W & Z \end{bmatrix}$, $\begin{bmatrix} P & S \\ Q & R \end{bmatrix}$ obtained from $\begin{bmatrix} \tilde{U} & \tilde{W} \\ \tilde{V} & \tilde{Z} \end{bmatrix}$, $\begin{bmatrix} \tilde{P} & \tilde{Q} \\ \tilde{S} & \tilde{R} \end{bmatrix}$ by partial integration.

The action of the differential operator $\begin{bmatrix} P & S \\ Q & R \end{bmatrix}$ on (23) thus results in the equivalent system of equations

$$\begin{aligned} -PGu_v + Sy_v &= v \\ -QG u_v + Ry_v &= 0 \end{aligned} \quad (33)$$

and, similarly, a combination of (31) and (24) yields

$$\begin{aligned} UH^T u_a - Vy_a &= p \\ WH^T u_a - Zy_a &= 0 \end{aligned} \quad (34)$$

It thus follows that $v(t) = -P(\frac{d}{dt})G(t)u_v(t) + S(\frac{d}{dt})y_v(t)$ and $p(t) = U(\frac{d}{dt})H^T(t)u_a(t) - V(\frac{d}{dt})y_a(t)$, implying that $\frac{d}{dt}p^T(t)v(t)$ appearing in the right-hand side of (14) can be expressed as a function Q in y_v , u_v , u_a , y_a and their time-derivatives (and implicitly of y , y , \dots , u , u , \dots). This shows that Definition 1 is consistent with the definition of the adjoint variational system Σ_a^* . Furthermore, as additional information we obtain from (33), (34) that input-output differential representations of Σ_a^v and Σ_a^* are given by

$$-Q\left(\frac{d}{dt}\right)G(t)u_v(t) + R\left(\frac{d}{dt}\right)y_v(t) = 0 \quad (35)$$

respectively

$$W\left(\frac{d}{dt}\right)H^T(t)u_a(t) - Z\left(\frac{d}{dt}\right)y_a(t) = 0. \quad (36)$$

Now compare (35) to (15), and write (15) in a more convenient notation as

$$\Sigma_{i/o}^v: D\left(\frac{d}{dt}\right)y_v(t) - N\left(\frac{d}{dt}\right)u_v(t) = 0. \quad (37)$$

The requirement, as alluded to before, that the set of solutions u_v , y_v generated by Σ_a^v equals the solution set of $\Sigma_{i/o}^v$ can thus be rephrased as the following.

Assumption 1: There exists an invertible matrix $E(\frac{d}{dt}) \in \mathcal{M}[D]^{p \times p}$ such that

$$\begin{aligned} E\left(\frac{d}{dt}\right)N\left(\frac{d}{dt}\right) &= Q\left(\frac{d}{dt}\right)G(t), \\ E\left(\frac{d}{dt}\right)D\left(\frac{d}{dt}\right) &= R\left(\frac{d}{dt}\right). \end{aligned} \quad (38)$$

From now on we will suppose throughout that Assumption 1 holds.

Our next objective is to give a procedure to compute an input-output differential representation of the adjoint variational system directly in terms of the input-output differential representation $\Sigma_{i/o}^v$ of the variational system (without going through the state space representation). Thus, let us consider $\Sigma_{i/o}^v$ as given by (15), denoted in shorthand notation by (37). Premultiply (37) by a p -dimensional row-vector $\xi^T(t)$ and again apply partial integration to

$$0 = \int_{t_1}^{t_2} \xi^T(t) \left[D\left(\frac{d}{dt}\right)y_v(t) - N\left(\frac{d}{dt}\right)u_v(t) \right] dt \quad (39)$$

to shift all differentiations in $y_v(t)$, $u_v(t)$ to differentiation in $\xi(t)$. This results in

$$\begin{aligned} 0 = \int_{t_1}^{t_2} y_v^T(t) \left[\tilde{D}\left(\frac{d}{dt}\right)\xi(t) \right] - u_v^T(t) \left[\tilde{N}\left(\frac{d}{dt}\right)\xi(t) \right] dt \\ + R(y, \dot{y}, \dots, u, \dot{u}, \dots, y_v, \dot{y}_v, \dots, u_v, \dot{u}_v, \dots, \xi, \dot{\xi}, \dots) \Big|_{t_1}^{t_2} \end{aligned} \quad (40)$$

for certain matrices \tilde{D} , \tilde{N} in $\mathcal{M}[D]$, and remainders R .

Comparing this to (17) (with Q playing the role of the remainders R) motivates the following definition of the adjoint variational system by $(\Sigma_{i/o}^a)$

$$\begin{aligned} u_a(t) &= D\left(\frac{d}{dt}\right)\xi(t) \\ y_a(t) &= N\left(\frac{d}{dt}\right)\xi(t). \end{aligned} \quad (41)$$

Note that (41) is an image representation, in contrast with the kernel representation (37). (For a linear time-invariant system (41) corresponds to a right factorization, while (37) corresponds to a left factorization of the transfer matrix.) Indeed, the (analytic) input-output behavior of (41) consists of all analytic time-functions $y_a(t)$, $u_a(t)$ satisfying (41) for some analytic function $\xi(t)$.

Theorem 1: The equations (41) are an image representation of the adjoint variational system defined in Definition 1.

Proof: We only have to show that R in (40) can also be expressed as a function of $y, \dot{y}, \dots, u, \dot{u}, \dots, y_v, \dot{y}_v, \dots, u_v, \dot{u}_v, \dots, y_a, \dot{y}_a, \dots, u_a, \dot{u}_a, \dots$.

From (25) we obtain

$$\begin{bmatrix} I \frac{d}{dt} - F & -G & 0 \\ -H & 0 & I_p \end{bmatrix} \begin{bmatrix} \tilde{U} & 0 \\ \tilde{V} & 0 \\ 0 & I_p \end{bmatrix} \begin{bmatrix} K & I_p \end{bmatrix} \quad (42)$$

for some K in $\mathcal{M}[D]$. Hence, from (32) and (42) we obtain

$$\begin{aligned} \begin{bmatrix} P & S \\ Q & R \end{bmatrix} \begin{bmatrix} I \frac{d}{dt} - F & -G & 0 \\ -H & 0 & I_p \end{bmatrix} \begin{bmatrix} \tilde{U} & 0 \\ \tilde{V} & 0 \\ 0 & I_p \end{bmatrix} \\ = \begin{bmatrix} -I_n & -PG & S \\ 0 & -QG & R \end{bmatrix} \begin{bmatrix} \tilde{U} & 0 \\ \tilde{V} & 0 \\ 0 & I_p \end{bmatrix} \\ = \begin{bmatrix} P & S \\ Q & R \end{bmatrix} \begin{bmatrix} I_n & 0 \\ -K & I_p \end{bmatrix}. \end{aligned} \quad (43)$$

Clearly the right-hand side of (43) is an invertible matrix, and thus "postmultiplication" of (43) by this inverse yields

$$-Q\left(\frac{d}{dt}\right)G(t)\bar{B}\left(\frac{d}{dt}\right) + R\left(\frac{d}{dt}\right)\bar{A}\left(\frac{d}{dt}\right) = I_p \quad (44)$$

for some matrices $\bar{A}(\frac{d}{dt})$, $\bar{B}(\frac{d}{dt})$ in $\mathcal{M}[D]$. Now recall that an input-output differential representation of Σ_s^v is given by (35), while also (38) holds. Thus there also exist matrices $A(\frac{d}{dt})$, $B(\frac{d}{dt})$ in $\mathcal{M}[D]$ such that

$$D\left(\frac{d}{dt}\right)A\left(\frac{d}{dt}\right) - N\left(\frac{d}{dt}\right)B\left(\frac{d}{dt}\right) = I_p. \quad (45)$$

Applying the partial integration procedure [sec (27)–(30)] to (45) yields

$$\tilde{A}\left(\frac{d}{dt}\right)D\left(\frac{d}{dt}\right) - \tilde{B}\left(\frac{d}{dt}\right)\tilde{N}\left(\frac{d}{dt}\right) = I_p \quad (46)$$

for certain matrices \tilde{A} , \tilde{B} in $\mathcal{M}[D]$, and with \tilde{D} , \tilde{N} as given by (41). Thus, by (41)

$$\begin{aligned} \xi(t) &= \tilde{A}\left(\frac{d}{dt}\right)\tilde{D}\left(\frac{d}{dt}\right)\xi(t) - \tilde{B}\left(\frac{d}{dt}\right)\tilde{N}\left(\frac{d}{dt}\right)\xi(t) \\ &= A\left(\frac{d}{dt}\right)u_a(t) - B\left(\frac{d}{dt}\right)y_a(t) \end{aligned}$$

showing that ξ can be expressed into y_a , u_a and their time-derivatives. ■

Remark 1 The above notions seem to be also useful for analyzing the controllability properties of an input-output differential system (11). Consider the variational and adjoint variational systems of (11) given by (37), respectively (41), where the entries of the matrix differential operators D , N , and \tilde{D} , \tilde{N} are seen as functions of $y, \dot{y}, \dots, u, \dot{u}, \dots$. Then one may construct matrix differential operators D_a , N_a (with entries depending on $y, \dot{y}, \dots, u, \dot{u}, \dots$) of maximal rank such that $D_a\tilde{N} - N_a\tilde{D} = 0$, implying that

$$D_a\left(\frac{d}{dt}\right)y_a(t) - N_a\left(\frac{d}{dt}\right)u_a(t) = 0 \quad (47)$$

for all trajectories u_a , y_a generated by (41). Integration by parts applied to the kernel representation (47) yields the "adjoint of the adjoint system," given in image representation

$$\begin{aligned} u_v(t) &= \tilde{D}_a\left(\frac{d}{dt}\right)\eta(t) \\ y_v(t) &= \tilde{N}_a\left(\frac{d}{dt}\right)\eta(t) \end{aligned} \quad (48)$$

where we have suggestively denoted the inputs and outputs by u_v and y_v , since we want to compare (48) to the variational system (37). Indeed, it is easily checked that all trajectories u_v, y_v generated by (48) satisfy (37), while equality of the behavior of (48) and (37) seems to correspond to some form of controllability of the original nonlinear system (11) [(48) defines the "controllable" part of the system]. This is an area for future research.

In the next section we will use the representation of the adjoint variational system $\Sigma_{i/o}^a$ as given by (41) in order to give a convenient characterization of Hamiltonian systems.

III. CHARACTERIZATION OF HAMILTONIAN SYSTEMS FROM THE INPUT-OUTPUT DIFFERENTIAL REPRESENTATION

We will derive in this section a complete characterization of Hamiltonian systems using the representation of the adjoint variational systems as in (41). In Section IV we will use an alternative approach based on the adjoint variational system Σ_s^a found in [5]. Note that for systems described as

$$\Sigma_{i/o} \quad F(y, \dot{y}, y, u, \dot{u}) = 0 \in \mathbb{R}, \quad u \in \mathbb{R}, y \in \mathbb{R} \quad (49)$$

the conditions under which (49) represents a Hamiltonian system have been found already in an earlier paper by Crouch and Lamnabhi [4], i.e.,

$$\begin{aligned} \text{i)} \quad & \frac{\partial F}{\partial u} = 0 \\ \text{ii)} \quad & \frac{\partial F}{\partial u} \left(\frac{d}{dt} \frac{\partial F}{\partial y} - \frac{\partial F}{\partial y} \right) = \frac{\partial F}{\partial y} \frac{d}{dt} \frac{\partial F}{\partial u} \end{aligned} \quad (50)$$

for every solution $(y, \dot{y}, y, u, \dot{u})$ satisfying (49).

Using the characterization of the adjoint variational system given in Section II we now obtain similar conditions for a general input-output differential representation (11).

Theorem 2 Consider a minimal input-output differential representation $\Sigma_{i/o}$ given by (11) with $p = m$, and its variational systems $\Sigma_{i/o}^v$ given by (37) satisfying Assumption 1. Compute the adjoint variational system $\Sigma_{i/o}^a$ given by (41). Then $\Sigma_{i/o}$ is an input-output representation of a Hamiltonian system if and only if

$$D \left(\frac{d}{dt} \right) \tilde{N} \left(\frac{d}{dt} \right) - N \left(\frac{d}{dt} \right) \tilde{D} \left(\frac{d}{dt} \right) = 0 \quad (51)$$

along every analytic solution $(y(t), u(t))$ of $\Sigma_{i/o}$.

Proof. Observe that (51) is equivalent to $\Sigma_{i/o}^v = \Sigma_{i/o}^a$, i.e., the input-output behavior defined by (37) is the same as the input-output behavior defined by (41). In the terminology of [5], [6] this means that every variational system Σ_s^v along an analytic solution of $\Sigma_{i/o}$ is self-adjoint. In [5, ch. 4] it is shown that Σ_s is Hamiltonian if and only if every variational system along trajectories resulting from piecewise constant inputs are self-adjoint. We finally note that by the Approximation Lemma [19, Lemma 1] the approximation of piecewise constant input functions by analytic input functions will result in state trajectories converging to the state trajectories corresponding to the piecewise constant input functions. ■

We will now work out in detail the self-adjointness condition (51) in case $N = 2$, i.e., we consider

$$\Sigma_{i/o}: \quad F(y, \dot{y}, \ddot{y}, u, \dot{u}) = 0 \in \mathbb{R}^m, \quad u, y \in \mathbb{R}^m \quad (52)$$

and

$$\Sigma_{i/o}^v: \quad Ay_v + B\dot{y}_v + C\ddot{y}_v + Du_v + E\dot{u}_v = 0, \quad u_v, y_v \in \mathbb{R}^m \quad (53)$$

where the (i, k) th elements of the $m \times m$ matrices A, B, C, D, E are given by

$$\begin{aligned} A_{ik} &= \frac{\partial F_i}{\partial y^k}, & B_{ik} &= \frac{\partial F_i}{\partial \dot{y}^k}, & C_{ik} &= \frac{\partial F_i}{\partial \ddot{y}^k}, \\ D_{ik} &= \frac{\partial F_i}{\partial u^k}, & E_{ik} &= \frac{\partial F_i}{\partial \dot{u}^k}. \end{aligned} \quad (54)$$

To compute the adjoint variational system $\Sigma_{i/o}^a$ two consecutive partial integrations yield (we are using summation convention)

$$\begin{aligned} 0 &= \int_{t_1}^{t_2} \xi^i [A_{ik} y_v^k + B_{ik} \dot{y}_v^k + C_{ik} \ddot{y}_v^k + D_{ik} u_v^k + E_{ik} \dot{u}_v^k] dt \\ &= \int_{t_1}^{t_2} [\xi^i A_{ik} y_v^k - (\xi^i B_{ik})^{(1)} y_v^k + (\xi^i C_{ik})^{(2)} y_v^k \\ &\quad + \xi^i D_{ik} u_v^k - (\xi^i E_{ik})^{(1)} u_v^k] dt \\ &\quad + [\xi^i B_{ik} y_v^k + \xi^i C_{ik} \dot{y}_v^k + \xi^i E_{ik} u_v^k - (\xi^i C_{ik})^{(1)} y_v^k]_{t_1}^{t_2} \end{aligned} \quad (55)$$

Thus the adjoint variational system is

$$\Sigma_{i/o}^a: \quad \begin{cases} \dot{u}_v^k = \xi^i A_{ik} - (\xi^i B_{ik})^{(1)} + (\xi^i C_{ik})^{(2)} \\ \dot{y}_v^k = -\xi^i D_{ik} + (\xi^i E_{ik})^{(1)} \end{cases} \quad k = 1, \dots, m \quad (56)$$

while Q is given by the terms between the brackets | | The self-adjointness condition (51) now becomes

$$\begin{aligned} \sum_{k=1}^m & -A_{jk} D_{ik} \xi^i + A_{jk} (E_{ik} \xi^i)^{(1)} - B_{jk} (D_{ik} \xi^i)^{(1)} \\ & + B_{jk} (E_{ik} \xi^i)^{(2)} - C_{jk} (D_{ik} \xi^i)^{(2)} + C_{jk} (E_{ik} \xi^i)^{(3)} \\ & + D_{jk} A_{ik} \xi^i - D_{jk} (B_{ik} \xi^i)^{(1)} + D_{jk} (C_{ik} \xi^i)^{(2)} \\ & + E_{jk} (A_{ik} \xi^i)^{(1)} - E_{jk} (B_{ik} \xi^i)^{(2)} + E_{jk} (C_{ik} \xi^i)^{(3)} = 0 \end{aligned} \quad (57)$$

along every solution $(u(t), y(t))$ of $\Sigma_{i/o}$. This gives the following conditions.

Terms with $(\xi^i)^{(3)}$.

$$\sum_{k=1}^m [C_{jk} E_{ik} + E_{jk} C_{ik}] = 0$$

Terms with $(\xi^i)^{(2)}$.

$$\begin{aligned} \sum_{k=1}^m & [B_{jk} E_{ik} - C_{jk} D_{ik} + 3C_{jk} E_{ik} + D_{jk} C_{ik} \\ & - E_{jk} B_{ik} + 3E_{jk} C_{ik}] = 0. \end{aligned}$$

Terms with $(\xi^i)^{(1)}$.

$$\begin{aligned} \sum_{k=1}^m & [A_{jk} E_{ik} - B_{jk} D_{ik} + 2B_{jk} \dot{C}_{ik} - 2C_{jk} \dot{D}_{ik} \\ & + 3C_{jk} \ddot{E}_{ik} - D_{jk} B_{ik} + 2D_{jk} \dot{C}_{ik} + E_{jk} A_{ik} \\ & - 2E_{jk} \dot{B}_{ik} + 3E_{jk} \ddot{C}_{ik}] = 0. \end{aligned}$$

Terms with ξ^1 :

$$\sum_{k=1}^m [-A_{jk} D_{ik} + A_{jk} \dot{E}_{ik} - j_k \dot{D}_{ik} + B_{jk} \ddot{E}_{ik} - C_{jk} \ddot{D}_{ik} + C_{jk} E_{ik}^{(3)} + D_{jk} A_{ik} - D_{jk} \dot{B}_{ik} + D_{jk} \ddot{C}_{ik} + E_{jk} \dot{A}_{ik} - E_{jk} \ddot{B}_{ik} + E_{jk} C_{ik}^{(3)}] = 0.$$

$i, j = 1, \dots, m$.

Hence, we have the following theorem.

Theorem 3: Consider a minimal system (52). Then it is Hamiltonian if and only if the following conditions

$$\begin{aligned} \mathcal{L}_1) & CE^T + EC^T = 0, \\ \mathcal{L}_2) & BE^T - CD^T + 3C\dot{E}^T + DC^T - EB^T + 3E\dot{C}^T = 0, \\ \mathcal{L}_3) & AE^T - BD^T + 2B\dot{E}^T - 2C\dot{D}^T + 3C\ddot{E}^T \\ & - DB^T + 2D\dot{S}^T + EA^T - 2E\dot{B}^T + 3E\ddot{C}^T = 0, \\ \mathcal{L}_4) & -AD^T + A\dot{E}^T - B\dot{D}^T + B\ddot{E}^T - CD^T \\ & + CE^{(3)T} + DA^T - D\dot{B}^T + D\ddot{C}^T \\ & + E\dot{A}^T - E\dot{B}^T + EC^{(3)T} = 0 \end{aligned} \quad (58)$$

hold along every solution $(u(t), y(t))$ of (52).

Elaboration of the Conditions \mathcal{L}_1 , \mathcal{L}_2 , \mathcal{L}_3 , and \mathcal{L}_4 in Special Cases

Case 1: Let us first consider the case $m = 1$, i.e., (47). If we assume $C' = \frac{\partial F}{\partial y} \neq 0$, then \mathcal{L}_1 yields

$$E = \frac{\partial F}{\partial \dot{u}} = 0$$

which is the condition (50)-i). Then \mathcal{L}_2 reduces to

$$-\frac{\partial F}{\partial y} \frac{\partial F}{\partial u} + \frac{\partial F}{\partial u} \frac{\partial F}{\partial y} = 0$$

which is automatically satisfied. Furthermore \mathcal{L}_3 amounts to

$$-\frac{\partial F}{\partial \ddot{y}} \frac{\partial F}{\partial u} - \frac{\partial F}{\partial u} \frac{\partial F}{\partial \ddot{y}} - 2 \frac{\partial F}{\partial y} \left(\frac{\partial F}{\partial u} \right)^{(1)} + 2 \frac{\partial F}{\partial u} \left(\frac{\partial F}{\partial \ddot{y}} \right)^{(1)} = 0$$

or

$$-\frac{\partial F}{\partial \ddot{y}} \frac{\partial F}{\partial u} - \frac{\partial F}{\partial \ddot{y}} \left(\frac{\partial F}{\partial u} \right)^{(1)} + \frac{\partial F}{\partial u} \left(\frac{\partial F}{\partial \ddot{y}} \right)^{(1)} = 0$$

which is the condition (50)-ii), and it is easily checked that \mathcal{L}_4 is precisely the time-derivative of (50)-ii).

Let us now derive a more explicit expression for the remainder Q in this particular case. Since $E = 0$, from (55)

$$Q = \xi B y_u + \xi C \dot{y}_u - (\xi C)^{(1)} y_u$$

and from (56)

$$\xi = -D^{-1} y_a.$$

If the conditions (50) hold we readily obtain

$$Q = [y_a, \dot{y}_a]^T \begin{bmatrix} 0 & C/D \\ -C/D & 0 \end{bmatrix} \begin{bmatrix} y_u \\ \dot{y}_u \end{bmatrix}. \quad (59)$$

Now let us take a closer look at (50). Writing out

$$\left(\frac{\partial F}{\partial \ddot{y}} \right)^{(1)} \frac{\partial^2 F}{\partial \ddot{y}^2} y^{(3)} + \frac{\partial^2 F}{\partial \ddot{y} \partial \ddot{y}} \ddot{y} + \frac{\partial^2 F}{\partial y \partial \ddot{y}} \dot{y} + \frac{\partial^2 F}{\partial u \partial \ddot{y}} \dot{u}$$

and

$$\left(\frac{\partial F}{\partial u} \right)^{(1)} = \frac{\partial^2 F}{\partial u \partial \ddot{y}} y^{(3)} + \frac{\partial^2 F}{\partial \ddot{y} \partial u} \ddot{y} + \frac{\partial^2 F}{\partial y \partial u} \dot{y} + \frac{\partial^2 F}{\partial u^2} \dot{u}$$

and collecting terms first with $y^{(3)}$ and then \dot{u} we obtain

$$\begin{cases} \frac{\partial F}{\partial u} \frac{\partial^2 F}{\partial \ddot{y}^2} - \frac{\partial^2 F}{\partial u \partial \ddot{y}} \frac{\partial F}{\partial \ddot{y}} = 0 \\ \frac{\partial F}{\partial u} \frac{\partial^2 F}{\partial u \partial \ddot{y}} - \frac{\partial^2 F}{\partial u^2} \frac{\partial F}{\partial \ddot{y}} = 0 \end{cases} \quad (60)$$

which means that the elements of the matrix given in (59) only depend on y, \dot{y} (the state of the system), and thus (59) defines a symplectic form ω on the state space $\{(y, \dot{y}) | y \in \mathbb{R}, \dot{y} \in \mathbb{R}\}$ with coordinate expression

$$\omega = \begin{pmatrix} 0 & \frac{\partial F}{\partial y} / \frac{\partial F}{\partial u} \\ -\frac{\partial F}{\partial y} / \frac{\partial F}{\partial u} & 0 \end{pmatrix}.$$

Case 2: Let us assume that the input-output representation $F_i, i = 1, \dots, m$, have the following particular form

$$F_i(y, \dot{y}, y, u, \dot{u}) = S_i(y, \dot{y}, y) - u_i. \quad (61)$$

This is the classical case (see [14]). Conditions $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$, and \mathcal{L}_4 reduce to

$$\begin{aligned} \mathcal{L}_1) & \text{ void} \\ \mathcal{L}_2) & \frac{\partial S}{\partial y} - \left(\frac{\partial S}{\partial \dot{y}} \right)^T = 0 \\ \mathcal{L}_3) & \frac{\partial S}{\partial y} + \left(\frac{\partial S}{\partial \dot{y}} \right)^T - 2 \frac{d}{dt} \left(\frac{\partial S}{\partial \dot{y}} \right)^T = 0 \\ \mathcal{L}_4) & \frac{\partial S}{\partial y} - \left(\frac{\partial S}{\partial \dot{y}} \right)^T + \frac{d}{dt} \left(\frac{\partial S}{\partial \dot{y}} \right)^T - \frac{d^2}{dt^2} \left(\frac{\partial S}{\partial \dot{y}} \right)^T = 0 \end{aligned}$$

at every point (y, \dot{y}, y) .

These conditions are precisely the conditions (2.1.17) in [14]. It follows that (61) represents the input-output behavior of a Hamiltonian system if and only if

$$S_i = R_{i,k}(y, \dot{y}) \ddot{y}^k + T_i(y, \dot{y}), \quad i = 1, \dots, m$$

where $R_{i,k}$ and T_i satisfy (2.2.9) in [14].

Case 3:

$$F_i(y, \dot{y}, \ddot{y}, u, \dot{u}) = S_i(y, \dot{y}) - u_i, \quad i = 1, \dots, m.$$

From the result of Santilli, it follows that

$$S_i = X_{i,k}(y) \dot{y}^k + Y_i(y), \quad i = 1, \dots, m$$

where $X_{i,k}$ and Y_i satisfy

$$\begin{aligned} X_{i,k} + X_{k,i} &= 0 \\ \frac{\partial X_{i,k}}{\partial y^k} + \frac{\partial X_{k,i}}{\partial y^i} + \frac{\partial X_{i,i}}{\partial y^k} &= 0 \\ \frac{\partial Y_i}{\partial y^k} - \frac{\partial Y_k}{\partial y^i} &= 0. \end{aligned}$$

Thus $X = (X_{i,j})$ defines the symplectic form ω . This also follows from (18), since $y_a^k = \xi^k$ and $Q = y_a^i X_{i,k} y_u^k$.

Example 1: As an example we consider the system given by the equations

$$\begin{aligned}\dot{q} &= u \\ \dot{p} &= \sin q + \alpha p \\ y &= p.\end{aligned}$$

The corresponding input-output differential equation is given by

$$F(y, \dot{y}, \ddot{y}, u, \dot{u}) = (\ddot{y} - \alpha \dot{y})^2 + (\dot{y} - \alpha y)^2 u^2 - u^2 \equiv 0. \quad (62)$$

Clearly, by inspecting the state representation of the system, we see that the system is Hamiltonian with the Hamiltonian function $H = \cos q + up$, if and only if $\alpha = 0$. This is not so evident, however, from the form of the input-output differential equation above. We shall employ condition (50) of Crouch and Lamnabhi [4] to show this. Condition i), $\partial F / \partial \dot{u} \equiv 0$, is trivially satisfied. We calculate the quantity

$$Z = \frac{\partial F}{\partial u} \left[\frac{d}{dt} \frac{\partial F}{\partial y} - \frac{\partial F}{\partial \dot{y}} \right] - \frac{\partial F}{\partial \ddot{y}} \frac{d}{dt} \frac{\partial F}{\partial u}$$

directly from F to obtain

$$\begin{aligned}Z &= (-4u)(1 - (\dot{y} - \alpha y)^2)[(y^{(3)} - \alpha \ddot{y}) \\ &\quad + \alpha(\ddot{y} - \alpha \dot{y}) - u^2(\dot{y} - \alpha y)] \\ &\quad - 4(\dot{y} - \alpha y)[2u(\dot{y} - \alpha y)(\dot{y} - \alpha \dot{y}) \\ &\quad - \dot{u}(1 - (\dot{y} - \alpha y)^2)].\end{aligned}$$

From $F(y, \dot{y}, \ddot{y}, u, \dot{u}) \equiv 0$ we obtain

$$(y^{(3)} - \alpha \ddot{y}) = \frac{u \dot{u}}{(\dot{y} - \alpha y)} - (\dot{y} - \alpha y)u^2 - \frac{u \dot{u}(\dot{y} - \alpha y)^2}{(\dot{y} - \alpha y)}$$

Substituting this expression into the expression for Z and again using the definition of F , we see that $Z \equiv 0$ if and only if

$$\alpha u(1 - (\dot{y} - \alpha y)^2)(y - \alpha \dot{y}) \equiv 0.$$

This equation can be satisfied for all $y(0)$, $\dot{y}(0)$, $u(0)$ only if $\alpha = 0$, so we conclude that the system (62) is Hamiltonian if and only if $\alpha = 0$, as we previously concluded.

It is also interesting to compare the adjoint system with equation (62). The variational system is given by

$$\begin{aligned}u_v((\dot{y} - \alpha y)^2 u - u) + y_v(-\alpha u^2(\dot{y} - \alpha y)) \\ + \dot{y}_v(u^2(\dot{y} - \alpha y) - \alpha(\dot{y} - \alpha \dot{y})) + \ddot{y}_v(y - \alpha(\dot{y})) = 0.\end{aligned} \quad (63)$$

Using the method of integration by parts introduced in (56), we see that the adjoint system is given by the equation

$$y_a = \xi((\dot{y} - \alpha y)^2 u - u)$$

$$u_a = \xi \alpha u^2(\dot{y} - \alpha y) + \xi(u^2(\dot{y} - \alpha y) - \alpha(\dot{y} - \alpha \dot{y})) - \xi(\dot{y} - \alpha \dot{y}).$$

Thus $\xi = y_a / ((\dot{y} - \alpha y)^2 u - u)$. An explicit expression for the adjoint variational system in input-output differential representation (34) is obtained by eliminating ξ from the equation above. Our theory guarantees that the resulting equation coincides with equation (63) if and only if $\alpha = 0$.

Example 2. In the case of a linear time-invariant system

$$Ay + B\dot{y} + C\ddot{y} + Du + E\dot{u} = 0, \quad u, y \in \mathbf{R}^m$$

with A, B, C, D, E constant $m \times m$ matrices, the conditions $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3, \mathcal{L}_4$ reduce to

$$CE^T + EC^T = 0$$

$$BE^T - CD^T + DC^T - EB^T = 0$$

$$AE^T - BD^T - DB^T + EA^T = 0$$

$$-AD^T + DA^T = 0,$$

which is equivalently to the equality

$$\begin{aligned}(D + Es)(A^T - B^T s + C^T s^2) \\ = (A + Bs + Cs^2)(D^T - E^T s)\end{aligned}$$

for all $s \in \mathbf{C}$. This last equality in turn is equivalent to the transfer matrix $G(s) = (A + Bs + Cs^2)^{-1}(D + Es)$ of the system to satisfy the condition (cf. [1], [15]) $G(s) = G^T(-s)$.

Example 3 It can be straightforwardly checked that the Euler-Lagrange equations with external forces u

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{y}} \right) - \frac{\partial L}{\partial y} = u, \quad u, y \in \mathbf{R}^m$$

for any Lagrangian function $L(y, \dot{y})$ satisfy (56) (see also Santilli [14]).

Consider now an Euler-Lagrange system

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_1} \right) - \frac{\partial L}{\partial q_1} = v_1, \quad v_1, q_1 \in \mathbf{R}$$

in interconnection with a static nonlinearity N (see van der Schaft [15] for details of interconnections), and assume for simplicity that $L(y, \dot{y}_1) = \frac{1}{2} m \dot{y}_1^2 - V(y_1)$, while the nonlinearity N is described by a differentiable function $y_2 = h(v_2)$. The interconnected system with outputs y_1, y_2 and inputs u_1, u_2 is given as (after elimination of v_1 and v_2)

$$\begin{aligned}m \ddot{y}_1 + \frac{\partial V}{\partial y_1}(y_1) - y_2 - u_1 &= 0 \\ h(y_1 + u_2) - y_2 &= 0.\end{aligned}$$

Computing the matrices A, B, C, D, E as in (51) yields

$$\begin{aligned}A &= \begin{bmatrix} \frac{\partial^2 V}{\partial y_1^2}(y_1) & -1 \\ \frac{\partial h}{\partial v_2}(y_1 + u_2) & -1 \end{bmatrix}, \\ C &= \begin{bmatrix} m & 0 \\ 0 & 0 \end{bmatrix}, \\ D &= \begin{bmatrix} -1 & 0 \\ 0 & \frac{\partial h}{\partial v_2}(y_1 + u_2) \end{bmatrix}\end{aligned}$$

while both B and E are zero. It is readily checked that conditions $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3, \mathcal{L}_4$ are satisfied, and thus the interconnected system is a Hamiltonian system for every scalar differentiable nonlinearity $y_2 = h(v_2)$.

In the multivariable case with $y_1, y_2, u_1, u_2, v_1, v_2 \in \mathbf{R}^m$ and $L(y_1, \dot{y}_1) = \frac{1}{2} \dot{y}_1^T M \dot{y}_1 - V(y_1)$, with $M = M^T > 0$,

it is straightforwardly checked that the interconnected system satisfies $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3, \mathcal{L}_4$ if and only if the nonlinearity $y_2 = h(v_2)$ satisfies the integrability condition $\frac{\partial h}{\partial v_2} = \left(\frac{\partial h}{\partial v_2}\right)^T$, and thus there exists (locally) a potential function $P(v_2)$ such that $y_2 = \frac{\partial P}{\partial v_2}(v_2)$.

IV. CHARACTERIZATION OF HAMILTONIAN SYSTEMS FROM THE STATE SPACE REPRESENTATION

In this section we derive another formulation of the criteria for self adjointness of the variational system (12) corresponding to an input-output differential representation (11). In this derivation, however, we work directly with the state space representations of the variational system (12) rather than the input-output differential representation (15) (which we did in the previous section). To do this we obtain a direct correspondence between the representations (12) and (15).

We first observe that the input-output map of the variational system Σ_v^u in equation (12) may be expressed in the form

$$y_v(t) = \int_0^t W_v(t, \sigma, u, x_0) u_v(\sigma) d\sigma \quad (64)$$

where we assume that $v(0) = 0$, and x_0 is the initial condition of the corresponding state space system (10). We also note that the input-output map of the adjoint variational system Σ_v^u in (13) is expressed in the form

$$y_u(t) = - \int_0^t W_v^T(t, \sigma, u, x_0) u_u(\sigma) d\sigma \quad (65)$$

where we assume that $p(0) = 0$. Thus (as formulated in [5], [6]) self-adjointness of the variational systems may be simply expressed as the statement

$$W_v(t, \sigma, u, x_0) = -W_v^T(\sigma, t, u, x_0) \quad (66)$$

for all $t \geq \sigma \geq 0$, all piecewise constant controls u , and one initial state x_0 . As we argued above, it is sufficient to check this identity for analytic controls u . Using the notation of (12), we may express the kernel W_v in the form

$$W_v(t, \sigma, x_0) = H(t)\Phi(t, \sigma)G(\sigma) (= W_v(t, \sigma))$$

where $\Phi(t, \sigma)$ is the transition matrix of $F(t)$. For analytic controls u , H , Φ , and G are all analytic in their arguments. Thus the self-adjointness criteria (66) may be expressed in the form

$$H(t)\Phi(t, \sigma)G(\sigma) = -G(t)^T \Phi(\sigma, t)^T H(t)^T \quad t \geq \sigma \geq 0. \quad (67)$$

Our first task is to give an equivalent formulation of the conditions (67) in terms of standard (time varying) linear system objects. We define the sequence of time varying $p \times n$ matrices $\Gamma_k(t)$ by setting

$$\begin{aligned} \Gamma_k^T(t) &= \left(I \frac{d}{dt} + F^T(t) \right)^k H^T(t); \\ \Gamma_0(t) &= H(t), \quad k \geq 0 \end{aligned} \quad (68)$$

and we define the sequence of time varying $n \times m$ matrices $\Sigma_k(t)$ by setting

$$\begin{aligned} \Sigma_k(t) &= (-1)^k \left(I \frac{d}{dt} - F(t) \right)^k G(t); \\ \Sigma_0(t) &= G(t), \quad k \geq 0. \end{aligned} \quad (69)$$

Noting that

$$\begin{aligned} \frac{\partial}{\partial t} \Phi(t, \sigma) &= F(t) \Phi(t, \sigma); \\ \frac{\partial}{\partial \sigma} \Phi(t, \sigma) &= -\Phi(t, \sigma) F(\sigma) \end{aligned}$$

we easily obtain

$$\frac{\partial^k}{\partial t^k} H(t) \Phi(t, \sigma) = \Gamma_k(t) \Phi(t, \sigma); \quad k \geq 0 \quad (70)$$

$$\frac{\partial^k}{\partial \sigma^k} \Phi(t, \sigma) G(\sigma) = (-1)^k \Phi(t, \sigma) \Sigma_k(\sigma); \quad k \geq 0. \quad (71)$$

Lemma 1. In the case of analytic data the identity (67) is equivalent to the sequence of identities

$$(-1)^{k+l+1} \Gamma_k(t) \Sigma_l(t) = \Sigma_k(t)^T \Gamma_l(t)^T; \quad k, l \geq 0. \quad (72)$$

Proof: By applying (70) and (71) to (67) we obtain

$$(-1)^l \Gamma_k(t) \Phi(t, \sigma) \Sigma_l(\sigma) = -(-1)^k \Sigma_k(t)^T \Phi(\sigma, t)^T \Gamma_l(\sigma)^T.$$

By setting $t = \sigma$ we obtain (72). Conversely, from (72) and analyticity we recover the identity

$$(-1)^{k+l+1} \Gamma_k(t) \Phi(t, \sigma) G(\sigma) = \Sigma_k(t)^T \Phi(\sigma, t)^T H(\sigma)^T.$$

For $k = 0$, this is the desired identity (67). ■

Our main interest is to show that we may replace the infinite set of conditions represented by (72), with a suitable finite subset. We require the following results.

Lemma 2. Consider a time varying state space system

$$\begin{aligned} \dot{x} &= F(t)x + G(t)u, \quad x(0) = 0, \quad x \in \mathbf{R}^l \\ y &= H(t)x. \end{aligned} \quad (73)$$

Then

$$\begin{aligned} \frac{d^p}{dt^p} y(t) &= \sum_{k=0}^{p-1} \frac{d^k}{dt^k} (\Gamma_{p-1-k} \Sigma_0 u)(t) \\ &\quad + \int_0^t \Gamma_p(t) \Phi(t, \sigma) G(\sigma) u(\sigma) d\sigma. \end{aligned} \quad (74)$$

Proof: Since $\Gamma_0 = H(t)$, the statement (74) is true for $p = 0$. Assume by induction that (74) is true. Then

$$\begin{aligned} \frac{d^{p+1}}{dt^{p+1}} y(t) &= \sum_{k=0}^{p-1} \frac{d^{k+1}}{dt^{k+1}} (\Gamma_{p-1-k} \Sigma_0 u)(t) + \Gamma_p(t) G(t) u(t) \\ &\quad + \int_0^t \Gamma_{p+1}(t) \Phi(t, \sigma) G(\sigma) u(\sigma) d\sigma \\ &= \sum_{k=1}^p \frac{d^k}{dt^k} (\Gamma_{p-k} \Sigma_0 u)(t) + \Gamma_p(t) \Sigma_0(t) u(t) \\ &\quad + \int_0^t \Gamma_{p+1}(t) \Phi(t, \sigma) G(\sigma) u(\sigma) d\sigma. \end{aligned}$$

This is, then, the identity (74) with p replaced by $p + 1$. ■

Lemma 3: The input-output map of the controllable time varying linear system (73) satisfies the system

$$\sum_{k=0}^N A_k(t) y^{(k)}(t) = \sum_{k=0}^{N-1} B_k(t) u^{(k)}(t) \quad (75)$$

if and only if

$$\sum_{k=0}^N A_k(t) \Gamma_k(t) = 0 \quad (76)$$

$$B_k(t) = + \sum_{q=k}^{N-1} A_{q+1}(t) \sum_{p=k}^q (\Gamma_{p-k}(t) \Sigma_0(t))^{(p-k)} \binom{p}{k}. \quad (77)$$

Proof: The input-output map of system (73) is given by

$$y(t) = \int_0^t H(t) \Phi(t, \sigma) G(\sigma) u(\sigma) d\sigma.$$

If this satisfies (75), we obtain the following expression using lemma (2)

$$\begin{aligned} & \sum_{p=1}^N A_p(t) \sum_{k=0}^{p-1} (\Gamma_{p-1-k} \Sigma_0 u)^{(k)}(t) \\ & + \sum_{p=0}^N A_p(t) \Gamma_p(t) \int_0^t \Phi(t, \sigma) G(\sigma) u(\sigma) d\sigma \\ & - \sum_{k=0}^{N-1} B_k(t) u^{(k)}(t) = 0. \end{aligned} \quad (78)$$

Thus

$$\begin{aligned} & \sum_{p=0}^N A_p(t) \Gamma_p(t) \int_0^t \Phi(t, \sigma) G(\sigma) u(\sigma) d\sigma = 0 \\ & \sum_{p=1}^N A_p(t) \sum_{k=0}^{p-1} \sum_{r=0}^k (\Gamma_{p-1-k} \Sigma_0)^{(k-r)}(t) u^{(r)}(t) \binom{k}{r} \\ & - \sum_{k=0}^{N-1} B_k(t) u^{(k)}(t) = 0. \end{aligned}$$

Controllability of system (73) and reordering summations now yields the desired identities (76) and (77). Conversely, the identities (76) and (77) yield (78) which by Lemma 2 ensures that the input-output map of (73) satisfies (75) as desired. ■

We wish to employ Lemma 3 in the context of the variational system Σ_s^v in (12) and the corresponding input-output representation $\Sigma_{i/o}^v$ in (15). However, (76) is not written in terms of purely input-output quantities which we require for our purposes. We therefore make the following observation.

If the system (73) is controllable, in the sense that

$$\text{rank} [\Sigma_0(t) | \Sigma_1(t) | \cdots | \Sigma_{N-1}(t)] = l \quad (79)$$

for all times t and all controls u , then the condition (76) may be replaced by the equivalent condition

$$\sum_{k=0}^N A_k(t) \Gamma_k(t) [\Sigma_0(t) | \Sigma_1(t) | \cdots | \Sigma_{N-1}(t)] = 0. \quad (80)$$

We may now state and prove our main result in this section.

Theorem 4: Consider a minimal input-output differential representation $\Sigma_{i/o}$, given by (11) with $p = m$, associated minimal state space representation Σ_s given by (11), and the associated variational system Σ_s^v . Then under the assumption (1) and the assumption that $\frac{\partial F}{\partial y(N)}$ is invertible for all analytic solutions (u, y) of (11), $\Sigma_{i/o}$ is an input-output representation of a Hamiltonian system Σ_s if and only if

$$(-1)^{k+l+1} \Gamma_k(t) \Sigma_l(t) = \Sigma_k(t)^T \Gamma_l(t)^T; \quad 0 \leq k, l \leq N \quad (81)$$

for all t and all analytic controls u .

Proof: As in [5] and [6], we know that Σ_s is Hamiltonian if and only if (66) holds for all analytic controls u . We have shown that these conditions are equivalent to the conditions (72), (Lemma 1). These conditions imply those of (81), hence establishing the necessity of the conditions (81). To prove necessity we argue as follows. We first compute the quantities Σ_k and Γ_k for the adjoint variational system Σ_s^a in (13). We make the substitutions

$$F \rightarrow -F^T; \quad G \rightarrow -H^T; \quad H \rightarrow G^T$$

in definitions (68) and (69). We obtain

$$\begin{aligned} \Sigma_k^a(t) &= (-1)^k \left(I \frac{d}{dt} + F(t)^T \right)^k (-H(t)^T) \\ &= (-1)^{k+1} \Gamma_k^v(t)^T = (-1)^{k+1} \Gamma_k(t)^T \end{aligned}$$

$$\begin{aligned} \Gamma_k^a(t) &= \left(I \frac{d}{dt} - F(t) \right)^k G(t) \\ &= (-1)^k \Sigma_k^v(t) = (-1)^k \Sigma_k(t). \end{aligned}$$

Thus

$$\Sigma_k^a(t) = (-1)^{k+1} \Gamma_k(t)^T; \quad \Gamma_k^a(t) = (-1)^k \Sigma_k(t)^T \quad (82)$$

We wish to generate conditions under which the input-output map (65) of Σ_s^a coincides with that of the input-output map (64) of Σ_s^v . By Assumption 1, the set of solutions (y_v, u_v) generated by Σ_s^v equals the set of solutions of $\Sigma_{i/o}^v$ represented by (15) or

$$\sum_{k=0}^N A_k(t) y_v^{(k)}(t) = \sum_{k=0}^{N-1} B_k(t) u_v^{(k)}(t). \quad (83)$$

(We also have that the set of solutions (y_v, u_v) of Σ_s^v corresponding to zero initial conditions $v(0) = 0$, is equal to the subspace through the origin of the set of solutions of $\Sigma_{i/o}^v$.) Hence, we may check self-adjointness of Σ_s^v , simply by checking that the input-output map (65) of Σ_s^a satisfies (83). By Lemma 3, however, the input-output map (65) satisfies (83) if and only if the (76) and (77) hold with Γ_k and Σ_k replaced by Γ_k^a and Σ_k^a , given in (82). Note that the controllability assumption required in Lemma 3 is translated into controllability of Σ_s^a , which is simply observability of Σ_s^v . Moreover, we may substitute the condition (76) by (80) as long as the condition (79) holds for Σ_s^a . Thus we obtain the

following conditions by substituting (82) into (80) and (77)

$$\sum_{k=0}^N A_k(t) \Sigma_k(t)^T (-1)^k [-\Gamma_0(t)^T | \Gamma_1(t)^T | \dots | \Gamma_{N-1}(t)^T] - \Gamma_2(t)^T \dots | (-1)^N \Gamma_{N-1}(t)^T] = 0$$

$$B_k(t) = \sum_{q=k}^{N-1} A_{q+1}(t) \sum_{p=k}^{N-1} (-1)^{p-k+1} (\Sigma_{p-k}^T(t) \Gamma_0(t)^T)^{(p-k)} \binom{p}{k}. \quad (84)$$

We now note that conditions (81) are indeed sufficient to ensure that (84) may be simply rewritten as (77) and (80). These are satisfied by virtue of the fact that by assumption the solutions (y_v, u_v) of Σ_s^v are solutions of $\Sigma_{i/o}^v$. It follows that we have shown sufficiency of the conditions (81), under the apparent further assumption that (79) holds for Σ_s^v .

To evaluate the further assumption it is useful to make an explicit construction, which is required later on. It is easily seen that we may rewrite the system $\Sigma_{i/o}^v$ given in (83) in the form

$$\sum_{k=0}^N (\hat{A}_k(t) y_v(t))^{(k)} = \sum_{k=0}^{N-1} (\hat{B}_k(t) u_v(t))^{(k)}$$

where $\hat{A}_N(t) = A_N(t) = \frac{\partial F}{\partial y^{(N)}}$. (Clearly the matrices \hat{A}_k and \hat{B}_k are related to the operators \hat{D} and \hat{N} defining the adjoint variational system in (41).) Under the assumption that $\frac{\partial F}{\partial y^{(N)}}$ is invertible we may rewrite (85) in the form

$$y_v^{(N)}(t) + \sum_{k=0}^{N-1} (\bar{A}_k(t) y_v(t))^{(k)} = \sum_{k=0}^{N-1} (\bar{B}_k(t) u_v(t))^{(k)}. \quad (85)$$

In this form we may write down the observable canonical form for the system (85), which will be a particular realization of Σ_s^v

$$\begin{pmatrix} -\bar{A}_{N-1}(t) & I & 0 & 0 \\ -\bar{A}_{N-2}(t) & 0 & I & 0 \\ \vdots & \vdots & \vdots & \vdots \\ -\bar{A}_0(t) & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \bar{B}_{N-1}(t) \\ \bar{B}_{N-2}(t) \\ \vdots \\ \bar{B}_0(t) \end{pmatrix} u_v(t) = y_v^{(N)}(t)$$

$$y(t) = (I, 0, \dots, 0)(x_1, x_2, \dots, x_N)^T. \quad (86)$$

Clearly this time varying system satisfies the condition

$$\text{Rank} [\Gamma_0(t)^T | \Gamma_1(t)^T | \dots | \Gamma_{N-1}(t)^T] = Np \quad (87)$$

for all t and all analytic controls u . Thus the corresponding adjoint system does satisfy the controllability condition (79). Note that the conditions (81) are independent of the particular realization of Σ_s^v which is chosen, and in particular there is no necessity for the chosen realization to be minimal (as a time varying system). ■

Finally in this section we point out the relationship between the conditions (81) and the conditions (49) derived in the previous section. Although the conditions (81) are apparently expressed in terms of the system Σ_s^v , or any other realization, we may interpret them directly in terms of $\Sigma_{i/o}^v$. In particular we may apply the conditions (81) to the realization (85) constructed above. The presentation of the resulting conditions on the matrices $A_k(t)$, $B_k(t)$ defining $\Sigma_{i/o}^v$, as in (83) turns out to be different, but equivalent, to the conditions obtained by applying Theorem 2 in the previous section. We demonstrate the conditions obtained in this section on the system (52). We write the corresponding variational system in the form of (53) but assume $C = I_m$, the identity matrix for ease of explanation. (The general conditions may be obtained by replacing A , B , D , and E by $C^{-1}A$, $C^{-1}B$, $C^{-1}D$, and $C^{-1}E$, respectively.) Now if the variational system is written as

$$A(t)y_v + B(t)\dot{y}_v + \ddot{y}_v = -E(t)\dot{u}_v - D(t)u_v$$

this may be rewritten in the form of (85) as follows

$$\ddot{y}_v + B(t)y_v + (A(t) - B(t))\dot{y}_v = (-E(t)u_v) + (-D(t) + \dot{E}(t))u_v.$$

Thus the corresponding observable canonical form (86) becomes

$$\frac{d}{dt} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -B(t) & I \\ \dot{B}(t) - A(t) & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} -E(t) \\ \dot{E}(t) - D(t) \end{pmatrix} u_v$$

$$y(t) = (I, 0)(x, x_2)^T. \quad (88)$$

The first condition in (81) is simply

$$-\Gamma_0(t)\Sigma_0(t) = \Sigma_0(t)^T\Gamma_0(t)^T$$

or simply

$$-H(t)G(t) = G(t)^T H(t)^T.$$

Substituting from (88) we obtain

$$E(t) = -E(t)^T.$$

Replacing E by $C^{-1}E$ we obtain

$$E(t)C(t)^T + C(t)E(t)^T = 0. \quad (89)$$

But this is simply \mathcal{L}_1 , in Theorem 3.

The next condition in (81) is simply

$$\Gamma_1(t)\Sigma_0(t) = \Sigma_1(t)^T\Gamma_0(t)^T$$

or simply $(\dot{H}(t) + H(t)F(t))G(t) = -(\dot{G}(t)^T - G(t)^T F(t)^T)H^T(t)$ or since $\dot{H}(t) \equiv 0$

$$H(t)F(t)G(t) = G(t)^T F(t)^T H(t)^T - \dot{G}(t)^T H(t)^T.$$

Substituting from (88) we obtain

$$\begin{aligned} [J, 0] \begin{bmatrix} -B & I \\ \dot{B} - A & 0 \end{bmatrix} \begin{bmatrix} -E \\ \dot{E} - D \end{bmatrix} \\ = [-E^T, \dot{E}^T - D^T] \begin{bmatrix} -B^T & \dot{B}^T - A^T \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix} \\ - [-\dot{E}^T, \ddot{E}^T - \dot{D}^T] \begin{bmatrix} I \\ 0 \end{bmatrix} \end{aligned}$$

which simply reduces to the condition

$$B(t)E(t) - E(t)^T B(t)^T + D(t)^T - D(t) + E(t) - 2E(t)^T = 0$$

Replacing E , B , and D by $C^{-1}E$, $C^{-1}B$, and $C^{-1}D$, we obtain

$$C^{-1}BC^{-1}E - E^T C^{-T} B^T C^{-T} + D^T C^{-T} - C^{-1}D + \overline{(C^{-1}E)} - 2\overline{(C^{-1}E)^T} = 0$$

However, from (89) we have

$$C^{-1}E = -E^T C^{-T}$$

By expanding the above expression and using this identity we obtain

$$\begin{aligned} & -C^{-1}BE^T C^{-T} + C^{-1}EB^T C^{-T} + D^T C^{-T} \\ & - C^{-1}D + C^{-1}C^T E^T C^{-T} + C^{-1}E \\ & - 2C^{-1}EC^T C^{-T} - 2E^T C^{-T} = 0 \end{aligned}$$

which may be reexpressed, using the invertibility of C , as

$$\begin{aligned} & -BE^T + EB^T + CD^T - DC^T + CE^T + EC^T \\ & - 2EC^T - 2CE^T = 0 \end{aligned}$$

By differentiating (89), however, we obtain

$$CE^T + EC^T = -EC^I - CE^I$$

We therefore obtain the expression

$$-BE^I + EB^T + CD^T - DC^T - 3EC^I - 3CE^I = 0$$

Now it is easily seen that this is just condition \mathcal{L}_2 in Theorem 3. Clearly, the remaining conditions \mathcal{L}_3 and \mathcal{L}_4 will be contained in the conditions (81) for $N = 2$.

We make two final comments on (81). Clearly, by inspecting (77) and (80), the number of conditions in (81) may be reduced to $N \geq k \geq 0$, $N - 1 \geq l \geq 0$, $k \geq l$. For $N = 2$ this results in five conditions, but we already know from Theorem 3 that in the case $N = 2$, there are only four independent conditions. Thus we expect even the reduced set of conditions (81) to include many redundancies. Furthermore, Theorem 4 requires the assumption that $\frac{\partial F}{\partial y^{(N)}}$ is invertible, which Theorem 2 does not.

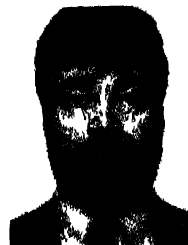
Conditions (81) do provide a satisfying generalization of the Brockett and Rahami result, discussed in the introduction. In particular, the condition that system (4) is Hamiltonian is simply given by the self-adjointness condition

$$C^T e^{A(t-\sigma)} B = -B^T e^{-A^T(t-\sigma)} C^T$$

By setting $\Gamma_k = C^T A^k$, $\Sigma_k = A^k B$, it is clear that self-adjointness is indeed equivalent to condition (81) for $N = n - 1$.

REFERENCES

- [1] R. W. Brockett and A. Rahami, "Lie algebras and linear differential equations," in *Ordinary Differential Equations*, L. Weiss, Ed., New York: Academic, 1972, pp. 379-386.
- [2] P. M. Cohn, *Free Rings and Their Relations*, 2nd ed., London: Academic, 1985.
- [3] J. M. Coron, "Linearized control systems and applications to smooth stabilization," *SIAM J. Contr. Opt.*, to be published.
- [4] P. E. Crouch and F. Lamnabhi Lagarrigue, "State space realization of nonlinear systems defined by input-output differential equations," in *Analysis and Optimization Systems*, Lect. Notes Contr. Inform. Sci., no. 111, 1988, pp. 138-149.
- [5] P. E. Crouch and A. J. van der Schaft, *Variational and Hamiltonian Control Systems*, Lect. Notes Contr. Inform. Sci., no. 101, 1987.
- [6] ———, "Hamiltonian and self adjoint control systems," *Syst. Contr. Lett.*, vol. 8, pp. 289-295, 1987.
- [7] S. Diop and M. Fliess, "Nonlinear observability, identifiability and persistent trajectories," in *Proc. 30th IEEE CDC*, Brighton, 1991, pp. 714-719.
- [8] M. Fliess, "Sur la réalisation des systèmes dynamiques bilinéaires," *C. R. Acad. Sci. Paris*, vol. A 277, pp. 243-247, 1973.
- [9] P. Ioannou and K. Tsakalis, *Linear Time Varying Systems: Control and Adaptation*, Englewood Cliffs, NJ: Prentice Hall, 1993.
- [10] A. Ilchmann, I. Nürnberger, and W. Schmale, "Time varying polynomial matrix systems," *Int. J. Contr.*, vol. 40, pp. 329-362, 1984.
- [11] A. Ilchmann, *Contributions to Time Varying Linear Control Systems*, Hamburg, Germany: Verlag an der Lottbek, 1989.
- [12] B. Jakubczyk, "Hamiltonian realizations of nonlinear systems," in *Theory and Applications of Nonlinear Control Systems*, C. I. Byrnes and A. Lindquist, Eds., Amsterdam, The Netherlands: North Holland, 1986, pp. 261-271.
- [13] T. Kailath, *Linear Systems*, Englewood Cliffs, NJ: Prentice Hall, 1980.
- [14] R. M. Santilli, *Foundations of Theoretical Mechanics I*, New York: Springer Verlag, 1978.
- [15] A. J. van der Schaft, "System theoretic descriptions of physical systems," *CWI Tract 3*, CWI Amsterdam, 1984.
- [16] ———, "Duality for linear systems: External and state space characterization of the adjoint system," in *Analysis of Controlled Dynamical Systems*, B. Bonnard, B. Bride, J. P. Gauthier, and I. Kupka, Eds., Birkhäuser, 1991, pp. 393-403.
- [17] F. D. Sontag, "Universal nonsingular controls," *Syst. Contr. Lett.*, vol. 19, pp. 221-224, 1992. Errata in vol. 20, p. 77, 1993.
- [18] H. J. Sussmann, "Single input observability of continuous time systems," *Math. Syst. Theory*, vol. 12, pp. 371-393, 1979.
- [19] ———, "Existence and uniqueness of minimal realizations of nonlinear systems," *Math. Syst. Theory*, vol. 10, pp. 263-284, 1977.
- [20] J. C. Willems and F. Lamnabhi Lagarrigue, "A note on the linearization around an equilibrium," *Appl. Math. Lett.*, vol. 4, pp. 29-34, 1991.



Peter E. Crouch (M 85, SM 91) was born in Newcastle upon Tyne, England, in 1951. He received the B.Sc. degree in engineering science and the M.Sc. degree in control theory from the University of Warwick, England, and the Ph.D. degree in applied sciences from Harvard University in 1977.

He taught in the Department of Engineering Sciences at Warwick University from 1977 to 1984. He has taught at Arizona State University since 1984 and is currently the Chair of Electrical Engineering and Director of the Center for Systems Science and

Engineering at Arizona State University. He has published 87 papers and one monograph entitled *Variational and Hamiltonian Control Systems* (with Arjan J. van der Schaft). His research interests lie in dynamical systems, nonlinear systems theory, and control theory, their applications to problems in electrical, mechanical and aerospace engineering and semiconductor manufacturing, and related problems in numerical simulation.

Dr. Crouch is Associate Editor at Large for IEEE TRANSACTIONS ON AUTOMATIC CONTROL and has served as an Associate Editor for the same journal and *Systems and Control Letters*. He is currently also Associate Editor for the *Journal of Dynamical and Control Systems*, *Mathematics of Control Signals and Systems* and the *Journal of Mathematical Control and Information*.



Françoise Lamnabhi-Lagarrigue received her Doctorat d'Etat in 1985 from the University Paris IX.

She is currently a Research Director at the National Scientific Research Center (CNRS) and works at the Laboratoire des Signaux et Systèmes, Gif sur Yvette, France. Her main research interests are in the analysis and control of nonlinear systems.

Dr Lamnabhi-Lagarrigue is an Associate Editor for *Systems and Control Letters* and *International Journal of Control*.



Arjan J. van der Schaft was born in Vlaardingen, The Netherlands, in February 1955. He received the University degree and Doctorate in mathematics from the University of Groningen in 1979 and 1983, respectively.

In 1982 he joined the Department of Applied Mathematics of the University of Twente, where he is presently an Associate Professor. He is (co-) author of the following books: *System Theoretic Descriptions of Physical Systems* (Amsterdam: CWI, 1984); *Variational and Hamiltonian Control Systems*

(with P. F. Crouch, Berlin: Springer, 1987); and *Nonlinear Dynamical Control Systems* (with H. Nijmeijer, New York: Springer, 1990). His research interests include the analysis and control of nonlinear and mechanical systems, and the mathematical modeling of physical systems.

Dr. Van der Schaft is Past Associate Editor of *Systems and Control Letters* and the *Journal of Nonlinear Science*, and presently serves as Associate Editor for the IFEL TRANSACTIONS ON AUTOMATIC CONTROL.

Language Convergence in Controlled Discrete-Event Systems

Yosef M. Willner and Michael Heymann

Abstract—Discrete-event systems modeled as state machines in the framework of Ramadge and Wonham are considered. In this paper, convergence of the language generated or marked by the system to a specified legal language is investigated. The convergence property is studied both with respect to open-loop (uncontrolled) systems and with respect to controlled systems. In the latter case, both questions of existence and synthesis of supervisors that force convergence are examined. Algorithms for verification of convergence and for synthesis of stabilizing supervisors are provided. Finally, the concept of asymptotic behavior of systems is defined, and questions related to this concept are investigated.

I. INTRODUCTION

IN the main paradigm of supervisory control of discrete-event systems (DES) [1]–[3], the system is modeled as a finite automaton that executes state transitions in response to a stream of events that occur nondeterministically and asynchronously. The supervisory control problem consists of synthesizing a supervisor whose task is to disable events in an orderly fashion, from a specified subset of controlled events, so as to confine the behavior of the supervised system to within a specified legal language. It is generally further required that the synthesized supervisor be minimally restrictive in the sense that, to achieve legal behavior, the smallest number of events is disabled.

There are situations, however, when legal supervision is either impossible or impractical. For example, when a system failure occurs, it may be driven to a state at which it is impossible to prevent the occurrence of illegal event sequences. Furthermore, even when the behavior of the system can be confined to within legal bounds, such a constraint may lead to the design of a supervisor that is overly restrictive. In such situations the question of convergence of the system to the legal behavior is of great interest. Intuitively, we say that the system converges to the legal behavior, if, after a finite and bounded number of transitions, it begins to behave legally. The supervisory control task is then to force such convergence to take place.

In [4] and [5] the problem of stabilization of discrete-event systems has been introduced, and the stabilization, or convergence, concept was presented in terms of legal

and illegal states of the system. In [4] legal behavior was specified by a set T of legal states. A system was then defined to be stable if, for any arbitrary initial state, it reaches a state in T after a finite and bounded number of state transitions and thereafter remains in T indefinitely. A system was called stabilizable if there exists a supervisor under which the supervised system is stable. Algorithms for synthesis of stabilizing supervisors were also presented. Optimal stabilizing supervisors were presented in [8].

It is quite clear that the legal behavior of a DES cannot always be specified by a set of (legal) states, which constitutes a static specification, but rather by a legal language which constitutes a dynamic specification. In the latter case the system is said to converge to legal behavior if, after a bounded number of illegal state transitions, the system produces only legal behavior, that is, strings of the specification language.

The concept of language convergence was first introduced by Kumar *et al.* (who called it language-stability) in [9], more recently in [16] and [18], and independently by Willner and Heymann in [15]. A similar concept was also studied in a somewhat different setting by Ozveren and Willsky in [19]. In [15] the convergence concept was formulated in terms of w -languages (i.e., languages that consist of infinite strings). In the present paper we formulate the problem in terms of languages over Σ^* and extend the discussion far beyond that of [15]. We provide algorithms for determining the convergence of a regular language to another regular language and for synthesizing supervisors that guarantee that convergence takes place whenever such supervisors exist.

We show that for suffix-closed legal languages, i.e., languages in which every suffix of a string is also a string of the language, the convergence problem can be resolved with algorithms of polynomial complexity. In fact, for suffix-closed languages the complexity is essentially of the same order as the static stabilization problem of [4]. When the specification language E is not suffix closed, the convergence problem is resolved by an algorithm of complexity that is polynomial in the size of the recognizer of the system language but exponential in the size of the recognizer of the specification language.

A further concept introduced in the present paper is the asymptotic behavior of a DES. The asymptotic behavior is the language that the system executes after performing an arbitrarily large number of state transitions. We discuss the problem of synthesizing a supervisor that guarantees the confinement of the asymptotic behavior of the supervised system within a given legal language. In the case of a

Manuscript received April 20, 1992; revised May 31, 1994. Recommended by Past Associate Editor, P. J. Ramadge. This work was supported in part by the Technion Fund for Promotion of Research.

Y. Willner is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

M. Heymann is with the Department of Computer Science, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 9408278.

suffix-closed specification, the inclusion of the asymptotic behavior in the specification language is equivalent to the convergence of the supervised system to the legal language. In the general case, inclusion of the asymptotic behavior in the specification language ensures the convergence of the behavior of the supervised system to the legal language. The inverse property, that language convergence implies confinement of the asymptotic behavior, is not in general true.

The paper is organized as follows. In Section II, a brief review of the basic elements of the framework of [1] is given, and the concept of language convergence is introduced. Algorithms for verifying language convergence for regular languages are presented in Section III. In Section IV we discuss the controlled convergence problem, i.e., the problem of existence and synthesis of supervisors that guarantee convergence. Problems associated with asymptotic behavior are considered in Section V. At various points of the paper our results are compared with the results obtained in [9] and [18].

II. PRELIMINARIES

We adopt the Ramadge-Wonham (RW) [1] discrete-event control formalism, that is, the DES under consideration is modeled as a deterministic automaton

$$P = (Q, \Sigma, \delta, q_0, Q_m) \quad (2.1)$$

where Q is a finite set of states, Σ is a finite set of event symbols, partitioned into two disjoint subsets Σ_c of controlled events and Σ_u of uncontrolled events. The map $\delta : \Sigma \times Q \rightarrow Q$ is a partial function called the state transition function, q_0 is the initial state, and $Q_m \subset Q$ is a subset of final (or marked) states.

Let Σ^* denote the set of all finite strings over Σ including the empty string ϵ . A subset $L \subset \Sigma^*$ is called a language over Σ .

Letting δ be extended in the standard way (see, e.g., [6]) to a function $\Sigma^* \times Q \rightarrow Q$, the language generated by P is defined as

$$\mathcal{L}(P) := \{t \in \Sigma^* \mid \delta(t, q_0)!\} \quad (2.2)$$

where the notation "!" means "is defined." The language marked by P is defined as

$$\mathcal{L}_m(P) := \{t \in \mathcal{L}(P) \mid \delta(t, q_0) \in Q_m\}. \quad (2.3)$$

A sequence of states $q_1, \dots, q_n \in R \subset Q$ is called a path of P contained in R (starting at q_1 and ending at q_n) if there exists a sequence of event symbols $\sigma_1, \dots, \sigma_{n-1}$ such that $q_{i+1} = \delta(\sigma_i, q_i)$ for each $i = 1, \dots, n-1$. When $q_n = q_1$ we call the path a cycle (in R). A subset $R \subset Q$ is called acyclic if P has no cycles in R . If P has no cycles in Q we also say that P is acyclic. We shall say that a state q' is accessible from a state q if there is a path of P starting at q and ending at q' . We shall denote by $\mathcal{A}(P, q)$ the set of all states of P that are accessible from q .

For a subset $\hat{Q} \subset Q$, we shall denote by $P_{\hat{Q}}$ the automaton obtained from P by replacing the initial state by the set \hat{Q} , that is

$$P_{\hat{Q}} := (Q, \Sigma, \delta, \hat{Q}, Q_m). \quad (2.4)$$

The language generated by $P_{\hat{Q}}$ is then defined as

$$\mathcal{L}(P_{\hat{Q}}) := \{t \in \Sigma^* \mid \delta(t, \hat{q})! \text{ for some } \hat{q} \in \hat{Q}\}. \quad (2.5)$$

The language marked by $P_{\hat{Q}}$ is then given by (2.3) upon replacing P by $P_{\hat{Q}}$. When \hat{Q} is a singleton $\{q\}$, we shall write P_q instead of $P_{\{q\}}$ and $P_{q_0} = P$.

If $P_1 = (Q_1, \Sigma, \delta_1, q_{10}, Q_{1m})$ and $P_2 = (Q_2, \Sigma, \delta_2, q_{20}, Q_{2m})$ are two automata over the event set Σ , we say that P_2 is a subautomaton of P_1 , denoted $P_2 \subset P_1$, if

$$a) \quad Q_2 \subset Q_1, \quad Q_{2m} \subset Q_{1m}, \quad q_{10} = q_{20}$$

$$b) \quad \delta_2(\sigma, q) = \delta_1(\sigma, q) \quad \forall \sigma \in \Sigma, q \in Q_2 \text{ s.t. } \delta_2(\sigma, q)!$$

Thus, a subautomaton is obtained from a given automaton by deleting from it transitions and/or states along with all the transitions incident on the deleted states. For a subset $\hat{Q} \subset Q$ we define the restriction of P to $Q - \hat{Q}$, denoted $P|_{\hat{Q}}$, as the subautomaton of P obtained by deleting from it all the states of \hat{Q} and the transitions incident on these states.

For given deterministic automata $P = (Q, \Sigma, \delta, q_0, Q_m)$ and $A = (X, \Sigma, \xi, x_0, X_m)$, the strict synchronous composition $P||A$ of P and A is defined as

$$P||A = (Q \times X, \Sigma, \alpha, (q_0, x_0), Q_m \times X_m)$$

where $\alpha : \Sigma \times Q \times X \rightarrow Q \times X$ is defined by

$$\alpha(\sigma, (q, x)) := \begin{cases} (\delta(\sigma, q), \xi(\sigma, x)) & \text{if } \delta(\sigma, q)! \text{ and } \xi(\sigma, x)! \\ \text{undefined} & \text{otherwise.} \end{cases}$$

It is well known that $\mathcal{L}(P||A) = \mathcal{L}(P) \cap \mathcal{L}(A)$ and $\mathcal{L}_m(P||A) = \mathcal{L}_m(P) \cap \mathcal{L}_m(A)$.

A supervisor for a DES P is a map $S : \mathcal{L}(P) \rightarrow 2^{\Sigma_c}$ such that for each $t \in \mathcal{L}(P)$, $S(t) \subset \Sigma_c$ is the set of controlled events that must be disabled next. The concurrent operation of the system P and the supervisor S , denoted by S/P , is called the closed-loop system. That is, a transition in S/P can take place whenever it can take place in P and is not disabled by S . The language $\mathcal{L}(S/P)$ generated by the closed-loop system is thus given recursively by

$$\begin{aligned} \epsilon &\in \mathcal{L}(S/P) \\ (\forall s \in \mathcal{L}(S/P)) s\sigma &\in \mathcal{L}(S/P) \Leftrightarrow s\sigma \in \mathcal{L}(P) \wedge \sigma \notin S(s). \end{aligned}$$

The language marked by S/P will be defined by

$$\mathcal{L}_m(S/P) = \mathcal{L}(S/P) \cap \mathcal{L}_m(P) \quad (2.6)$$

and consists of all the strings marked by P that are not disabled by S .

We shall assume that the system P under consideration is trim, i.e., every state $q \in Q$ is accessible from q_0 and $\mathcal{L}(P) = \mathcal{L}_m(P)$. A supervisor S is then called nonblocking if $\mathcal{L}(S/P) = \mathcal{L}_m(S/P)$.

For strings $s, t \in \Sigma^*$ we let $s \cdot t$ denote their concatenation. A string s is a prefix of t , denoted $s \leq t$, if $t = sw$ for some $w \in \Sigma^*$. If $w \neq \epsilon$, s is said to be a proper prefix of t , denoted $s < t$. The prefix closure \bar{L} of a language $L \subset \Sigma^*$ is the set of all prefixes of strings in L and L is prefix closed if $\bar{L} = L$. The length of a string t is denoted $|t|$ and the prefix of t of length i is denoted $pr_i(t)$ ($pr_0(t) = \epsilon$).

For a string $t \in \Sigma^*$ we denote by $\text{suf}_i(t)$ the string obtained by deleting from t its first i symbols ($\text{suf}_0(t) = t$ and $\text{suf}_{|t|}(t) = \epsilon$). For a language $L \subset \Sigma^*$, the suffix closure of L is the subset of all suffixes of strings in L , that is

$$\text{suf}(L) = \{\text{suf}_i(t) \mid \forall t \in L, \forall i \leq |t|\}. \quad (2.7)$$

A language L is suffix closed if $\text{suf}(L) = L$. The following properties of the suffix operator are readily verified:

- The suffix operator is monotone, i.e.,

$$L_1 \subset L_2 \Rightarrow \text{suf}(L_1) \subset \text{suf}(L_2). \quad (2.8)$$

- For a family of languages $\{L_\alpha \subset \Sigma^*\}$

$$\text{suf}\left(\bigcup L_\alpha\right) = \bigcup \text{suf}(L_\alpha), \quad (2.9)$$

$$\text{suf}\left(\bigcup L_\alpha\right) \subset \bigcup \text{suf}(L_\alpha) \quad (2.10)$$

with equality in (2.10) if for each α , L_α is suffix closed.

Let $L \subset \Sigma^*$ be a regular language and let $P = (Q, \Sigma, \delta, q_0, Q_m)$ be a (trim) recognizer for L , i.e., $\mathcal{L}_m(P) = L$ and $\mathcal{L}(P) = \overline{\mathcal{L}_m(P)}$. A recognizer for $\text{suf}(L)$ is then given by $P_Q = (Q, \Sigma, \delta, Q, Q_m)$, that is, $\mathcal{L}_m(P_Q) = \text{suf}(L)$ (with $\mathcal{L}_m(P_Q)$ as defined following (2.5)). From (2.7), $L \subset \text{suf}(L)$, so that L is suffix closed if and only if $\mathcal{L}_m(P_Q) \subset \mathcal{L}_m(P)$. This inclusion is equivalent to the condition that $\mathcal{L}_m(P_Q) \cap (\Sigma^* - \mathcal{L}_m(P)) = \emptyset$, which can be verified by an algorithm with complexity of $O(|Q|^2)$ (see e.g., [10, Proposition 2.3]).

We conclude this section with the definition of language convergence that will be needed in the sequel. Let $L, M \subset \Sigma^*$ be two languages such that $M \neq \emptyset$.

Definition 2.1: The language L is said to converge asymptotically to M , denoted $M \leftarrow L$, if for each $t \in L$ there exists an integer $i \geq 0$ such that $\text{suf}_i(t) \in M$.

Remark 2.1: It is worth noting that the asymptotic convergence $M \leftarrow L$ is equivalent to $L \subset \Sigma^* \cdot M$ (where $\Sigma^* \cdot M$ denotes the language concatenation, i.e., $s \in \Sigma^* \cdot M$ if and only if $s = r \cdot t$ where $r \in \Sigma^*$ and $t \in M$). Thus, the asymptotic convergence problem can be investigated as a special case of the language confinement problem that was investigated in detail by RW in [1].

Definition 2.2: The language L is said to converge finitely (or, simply, to converge) to M , denoted $M \Leftarrow L$, if there exists a finite integer $k \geq 0$ such that for each $t \in L$ there exists $i, i \leq k$, for which $\text{suf}_i(t) \in M$. In that case the least k for which the above holds is called the convergence time of L to M and is denoted $e_M(L)$.

While finite convergence obviously implies asymptotic convergence, the converse is not generally true as illustrated by the simple example where $L = \alpha^* \beta$ and $M = \{\beta\}$.

It is easily verified that the class of languages that converge asymptotically to a given language is closed under arbitrary intersections and arbitrary unions while the class of languages that converge finitely to a given language is closed under arbitrary intersections but only under finite unions. Finally,

if $L_1, L_2, M \subset \Sigma^*$ are languages such that $L_1 \subset L_2$ and $M \neq \emptyset$, then

$$(M \leftarrow L_2) \Rightarrow (M \leftarrow L_1) \quad (2.11)$$

and

$$(M \Leftarrow L_2) \Rightarrow (M \Leftarrow L_1). \quad (2.12)$$

III. FINITE LANGUAGE CONVERGENCE

In the present section we shall develop necessary and sufficient conditions for finite language convergence. Furthermore, we develop algorithms for testing whether $M \Leftarrow L$ for regular languages $L, M \subset \Sigma^*, M \neq \emptyset$. For a variety of interesting special cases, algorithms of polynomial time complexity are constructed. In the most general case, however, our algorithm is of polynomial complexity in the size of the recognizer for L but exponential in the size of the recognizer for M .

Let $L, M \subset \Sigma^*, M \neq \emptyset$ be two regular languages, let $P = (Q, \Sigma, \delta, q_0, Q_m)$ and $A = (X, \Sigma, \xi, x_0, X_m)$ be deterministic trim recognizers for L and M , respectively, so that $\mathcal{L}_m(P) = L$, $\mathcal{L}_m(A) = M$, $\overline{\mathcal{L}_m(P)} = \mathcal{L}(P)$ and $\overline{\mathcal{L}_m(A)} = \mathcal{L}(A)$. Let T_q be the subset of states in Q given by

$$T_q = \{q \in Q \mid \mathcal{L}_m(P_q) \subset \text{suf}(M)\} \quad (3.1)$$

(where, as defined earlier, P_q is the automaton P initialized at q).

The following is a necessary condition for (finite-) convergence of L to M .

Proposition 3.1 Let $L, M \subset \Sigma^*$ be the languages recognized by the automata P and A , respectively. Then $M \Leftarrow L$ only if the set $Q - T_q$ is acyclic (in P).

Proof: Suppose that P has a cycle C in $Q - T_q$. Since P is trim, every state of P is accessible from q_0 and there is a sequence of strings $\{t_i\}$, $t_i \in \mathcal{L}(P) = \overline{L}$, $t_1 < t_2 < t_3 \dots$ such that for each i , $\delta(t_i, q_0) \in C$. Assume that $M \Leftarrow L$ with convergence time $k (= e_M(L))$. Let t_j be a string of the above sequence such that $n := |t_j| > k$. Then $q = \delta(t_j, q_0) \in C$ and there exists $s \in \mathcal{L}_m(P_q)$ such that $s \notin \text{suf}(M)$ (since the states of C do not belong to T_q). But $t_j \cdot s \in \mathcal{L}_m(P) = L$ and by the convergence of L to M , there exists $l \leq k$ such that $\text{suf}_l(t_j \cdot s) \in M$. Since $n > l$, it follows that $s = \text{suf}_n(t_j \cdot s) = \text{suf}_{n-l}(\text{suf}_l(t_j \cdot s)) \in \text{suf}(M)$. This is a contradiction, and the proof is complete. \square

Next, we present a sufficient condition for convergence of L to M . Let T be the set of all states in Q defined by

$$T := \{q \in Q \mid \mathcal{L}_m(P_q) \subset M\} \quad (3.2)$$

and assume that $q_0 \notin T$, for otherwise $L \subset M$ and convergence is trivial. Next, let R be defined as the set of all states of $Q - T$ that are accessible from q_0 through a path that does not intersect T , that is

$$R := \{q \in Q \mid \exists t \in \overline{L}, \delta(t, q_0) = q \text{ and } (\forall i \leq |t|) \delta(pr_i(t), q_0) \in Q - T\}. \quad (3.3)$$

The set R can conveniently be computed as follows. Consider the automaton $P|T$ (obtained from P by deleting from it all the states of T and the transitions incident on these states).

Then R is the set of all states accessible in $P|T$ from the initial state q_0 , that is

$$R = \mathcal{A}(P|T, q_0) \quad (3.4)$$

(of course, if $q_0 \in T$ then $L \subset M$ and $R = \phi$).

Proposition 3.2: If R is acyclic and $R \cap Q_m = \phi$, then $M \Leftarrow L$.

Proof: Assume first that $T = \phi$. Then (since P is trim) $R = Q$ so that $R \cap Q_m = Q_m$. If $R \cap Q_m = \phi$, then $Q_m = \phi$ and hence $L = \mathcal{L}_m(P) = \phi$ and $M \Leftarrow L$ holds trivially. Assume now that $T \neq \phi$. If $R \cap Q_m = \phi$, then for each $t \in L = \mathcal{L}_m(P)$ there must exist $i, i \leq |t|$, such that $q = \delta(pr_i(t), q_0) \in T$. Now the acyclicity of R implies that this i must satisfy the condition that $i \leq |R|$. Hence we conclude that for each $t \in L$ there exists $i \leq |R|$ such that $q = \delta(pr_i(t), q_0) \in T$, implying that $\text{suf}_i(t) \in \mathcal{L}_m(P_q) \subset M$. It follows that $M \Leftarrow L$ and $c_M(L) \leq |R|$. \square

Proposition 3.2 suggests the following algorithm for determining the convergence of L to M .

Algorithm 3.1:

- 1) Compute the set T of states defined by (3.2). If $q_0 \in T$ then $L \subset M$. Convergence holds trivially. Halt.
- 2) Compute the set R as defined by (3.4).
- 3) Check (e.g., by the algorithm of [7, p. 64]) whether P has cycles in R . If the answer is yes, the algorithm is inconclusive. Halt.
- 4) If $R \cap Q_m = \phi$, then $M \Leftarrow L$. Halt.

The computation of R (which is a standard accessibility computation) and the algorithm for testing acyclicity of R as given in [7] are both of complexity¹ $O(|Q|)$. As regards the computation of T we proceed as follows.

A state $q \in Q$ is in T if and only if $\mathcal{L}_m(P_q) \subset \mathcal{L}_m(A)$ where $A = (X, \Sigma, \xi, x_0, X_m)$ is a deterministic recognizer for M ($\mathcal{L}_m(A) = M$). Equivalently, $q \in T$ if and only if $\mathcal{L}_m(P_q) \cap \mathcal{L}_m(A) = \phi$. A deterministic recognizer for $\Sigma^* - \mathcal{L}_m(A)$ is given by the automaton $\hat{A} = (\hat{X}, \Sigma, \hat{\xi}, x_0, \hat{X}_m)$, where $\hat{X} = X \cup \{d\}$, d being an additional "dump" state, $\hat{X}_m = \{d\} \cup (X - X_m)$, and $\hat{\xi} : \Sigma \times \hat{X} \rightarrow \hat{X}$ is given by

$$\hat{\xi}(\sigma, \hat{x}) = \begin{cases} \xi(\sigma, \hat{x}) & \text{if } \hat{x} \neq d \text{ and } \xi(\sigma, x)! \\ d & \text{otherwise.} \end{cases}$$

The construction of the automaton \hat{A} is of complexity $O(|X|)$.

For a state $q \in Q$, a recognizer for $\mathcal{L}_m(P_q) \cap (\Sigma^* - \mathcal{L}_m(A))$ is given by the automaton $P_q||\hat{A}$ (see Section II) and we have that $\mathcal{L}_m(P_q) \subset \mathcal{L}_m(A)$, or equivalently $q \in T$, if and only if $\mathcal{L}_m(P_q||\hat{A}) = \phi$. Thus, to compute T , we construct the automaton

$$P_Q||\hat{A} = (Q \times \hat{X}, \Sigma, \alpha, \{q, x_0\} | q \in Q, Q_m \times \hat{X}_m)$$

where $\alpha(\sigma, (q, \hat{x})) = (\delta(\sigma, q), \hat{\xi}(\sigma, \hat{x}))$ is defined if and only if $\delta(\sigma, q)!$ and $\hat{\xi}(\sigma, \hat{x})!$. Next, we let $C(Q_m \times \hat{X}_m)$ be the set

of all states $(q, \hat{x}) \in Q \times \hat{X}$ from which the set $Q_m \times \hat{X}_m$ is accessible in $P_Q||\hat{A}$, that is

$$C(Q_m \times \hat{X}_m) := \{(q, \hat{x}) \in Q \times \hat{X} | \exists t \in \Sigma^*, \alpha(t, (q, \hat{x})) \in Q_m \times \hat{X}_m\}.$$

A state q is thus in T if and only if $(q, x_0) \notin C(Q_m \times \hat{X}_m)$, whence T is given by

$$T = Q - \{q \in Q | (q, x_0) \in C(Q_m \times \hat{X}_m)\}.$$

The computation of $C(Q_m \times \hat{X}_m)$ is of complexity $O(|Q||X|)$ and hence the overall complexity of Algorithm 3.1 is $O(|Q||X|)$.

An interesting case of the convergence problem is when the language M includes the empty string ϵ . This implies that all deadlocking behavior of L is regarded as convergent. Specifically, since for each string $t \in L$, $\text{suf}_{|t|}(t) = \epsilon \in M$, it follows that if all strings of L are of bounded length, then L converges to any language M satisfying $\epsilon \in M$. A further observation is the following. When $\epsilon \in M$, then the acyclicity of R (as defined in (3.3) or (3.4)) implies convergence of L to M . Indeed, for $t \in L$ such that $|t| \leq |R|$, $\text{suf}_{|t|}(t) = \epsilon \in M$, and for $t \in L$ such that $|t| > |R|$, the acyclicity of R implies that there exists $i \leq |R|$ such that $\text{suf}_i(t) \in M$. Thus we have the following.

Proposition 3.3: Let $L, M \subset \Sigma^*$ be regular languages recognized by automata P and A , respectively, and let R be the set defined by (3.4). If $\epsilon \in M$ and R is acyclic, then $M \Leftarrow L$.

A special case of languages M that include the empty string are suffix-closed languages. Such language specifications correspond to cases when one is interested in the eventual correctness of the logical behavior of the process P but does not insist on specific initialization. Thus, one is satisfied with the fact that the behavior of P "merges" with that of a given specification language rather than performing complete strings of the specification and one wishes to find conditions for $\text{suf}(M) \Leftarrow \mathcal{L}_m(P) = L$.

Let M be suffix closed. Then $\epsilon \in M$, and it can be further noted that the sets T_s as defined in (3.1) and T as defined in (3.2) coincide. Proposition 3.1 then states that if $M \Leftarrow L$ then $Q - T$ is acyclic, and since $R \subset Q - T$, so is also R . Combining this fact with Proposition 3.3 gives the following interesting theorem.

Theorem 3.1: Let $L, M \subset \Sigma^*$ be (regular) languages recognized by automata P and A , respectively. Assume that M is suffix closed. Then $M \Leftarrow L$ if and only if $Q - T$ is acyclic.

Theorem 3.1 provides the theoretical foundation for the following algorithm that determines the convergence of L to M in case L and M are regular languages and M is suffix closed.

Algorithm 3.2:

- 1) Compute the set T of states defined by (3.2). If $q_0 \in T$ then $L \subset M$. Convergence holds trivially. Halt.
- 2) Check (e.g., by the algorithm of [7, p. 64]) whether P has cycles in $Q - T$.
- 3) L converges to M if and only if the answer to 2) is negative. Halt.

¹In the present paper we shall assume $|\Sigma| = O(1)$ and we give the complexity bounds only in terms of the number of states.

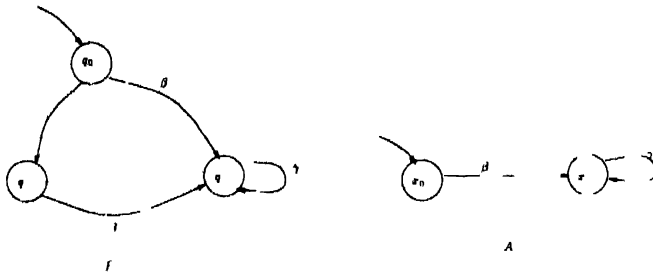


Fig 1

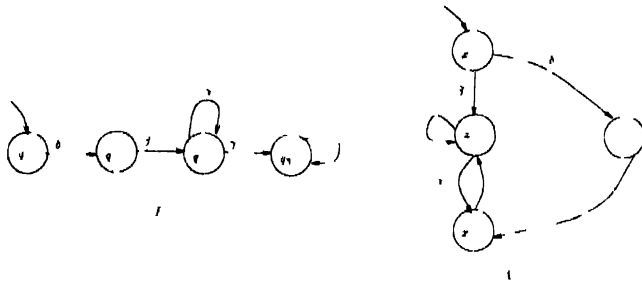


Fig 2

The complexity of Algorithm 3.2 is $O(|Q||\Sigma|)$ (see the analysis of complexity of Algorithm 3.1 for details).

We turn next to the problem of language convergence in the case when M is a general (not necessarily suffix-closed) regular language. The conditions of Proposition 3.2, and in the case when $\epsilon \in M$ of Proposition 3.3, are sufficient for convergence and, in view of their relative efficiency, should be tested first. When the corresponding algorithms (Algorithms 3.1 and 3.2, respectively) are inconclusive, one must resort to a more elaborate examination. As we shall see later, the algorithms in the general case are no longer of polynomial complexity.

Let us first examine two examples that will prove helpful in obtaining some intuitive insight. The first example shows that while the acyclicity of R is sufficient for convergence of L to M in case $\epsilon \in M$ (Proposition 3.3), it is not necessary when M is not suffix-closed.

Example 3.1 Consider the automata P and A as given in Fig. 1, where all the states of both P and A are final states.

Thus, $\mathcal{L}_m(P) = \mathcal{L}(P)$ and $\mathcal{L}_m(A) = \mathcal{L}(A)$ and, obviously, $\mathcal{L}(A) \subsetneq \mathcal{L}(P)$ (Here and in subsequent figures the entrance arrows indicate the initial states.) Observe that $T = \{q_1\}$, i.e., q_1 is the only state $q \in Q$ such that $\mathcal{L}(P_q) \subset \mathcal{L}(A)$ and the set $R = \{q_0, q_2\}$ is not acyclic. \square

Example 3.2 Here P and A are given by the automata of Fig. 2 and again all states (of P and A) are final states.

To verify that $\mathcal{L}(A) \subsetneq \mathcal{L}(P)$ note that

$$\mathcal{L}(P) = \overline{\delta\beta\alpha^*\gamma^*} = \delta\beta(\alpha\alpha)^*\gamma^* + \delta\beta\alpha(\alpha\alpha)^*\gamma^*$$



Fig 3

and that both $\overline{\beta(\alpha\alpha)^*\gamma^*}$ and $\overline{\delta\beta\alpha(\alpha\alpha)^*\gamma^*}$ are sublanguages of $\mathcal{L}(A)$.

Next note that

$$\mathcal{L}(P_{q_2}) = \alpha^*\gamma^* = (\alpha\alpha)^*\gamma^* + \alpha(\alpha\alpha)^*\gamma^*$$

and that $(\alpha\alpha)^*\gamma^* \subset \mathcal{L}(A_{r_1})$ and $\alpha(\alpha\alpha)^*\gamma^* \subset \mathcal{L}(A_{x_2})$.

Thus

$$\mathcal{L}(P_{q_2}) \subset \mathcal{L}(A_{r_1}) \cup \mathcal{L}(A_{x_2}) = \mathcal{L}(A_{\{r_1, x_2\}})$$

Note further that if $\delta(t, q_0) = q_2$ for $t \in \Sigma^*$, then $\xi(t, r_1) \in \{x_1, x_2\}$. \square

In Example 3.2 the language $L = \mathcal{L}(P)$ (that converges to $M = \mathcal{L}(A)$) satisfies the condition that the system P , starting at q_0 , reaches a state q such that

$$\mathcal{L}(P_q) \subset \mathcal{L}(A_\chi) \quad (3.5)$$

for some subset $\chi \subset \Sigma$. We shall see below that this is actually a necessary condition for convergence of L to M . Condition (3.5) is of course not sufficient for convergence as can be seen from the example in Fig. 3.

Here $\mathcal{L}(P)$ does not converge to $\mathcal{L}(A)$ although $\mathcal{L}(P_q) \subset \mathcal{L}(A_\chi)$.

Next we present an algorithm that tests the convergence of L to M for general regular languages L and M . The algorithm is based on the construction of a new recognizer for the language L in which the necessary condition that the automaton reaches a state that satisfies condition (3.5) is also sufficient. The complexity of the proposed algorithm is exponential in $|\Sigma|$, the dimension of the state set of A , but it is polynomial in $|Q|$, the dimension of the state set of P .

Theorem 3.2 Let $P = (Q, \Sigma, \delta, q_0, Q_m)$ and $A = (\Sigma, \Sigma, \xi, x_0, X_m)$ be given automata. The following algorithm verifies the convergence $\mathcal{L}_m(A) \subseteq \mathcal{L}_m(P)$.

Algorithm 3.3

- 1) Construct a deterministic automaton $B = (V, \Sigma, \alpha, v_0, V_m)$, where $V = Q \times 2^\Sigma$, $v_0 = (q_0, \{x_0\})$, $V_m = Q_m \times \Sigma$, and where $\alpha: \Sigma \times V \rightarrow V$ is defined as follows. Let $\chi \in 2^\Sigma$ be any element. Then see (3.6) at the bottom of the page. We assume that B is trim. Otherwise construct the maximal trim subautomaton of B .
- 2) Test whether there exists in B a state $v = (q, \chi) \in V$ such that $q \in Q_m$ and $\chi \cap X_m = \emptyset$. If such a state exists, $\mathcal{L}_m(A) \not\subseteq \mathcal{L}_m(P)$. Halt.
- 3) Compute the set $T \subset V$ defined by

$$T = \{v = (q, \chi) \in V \mid \mathcal{L}_m(P_q) \subset \mathcal{L}_m(A_\chi)\} \quad (3.7)$$

$$\alpha(\sigma, (q, \chi)) = \begin{cases} (\delta(\sigma, q), \chi') & \text{where } \chi' = \{x_0\} \cup \{x' \in \chi \mid (\exists x \in \chi) \xi(\sigma, x) = x'\} \\ \text{undefined} & \text{otherwise} \end{cases} \quad (3.6)$$

- 4) Test whether $V - \tilde{T}$ is acyclic (in B) (use, e.g., the algorithm of [7]).
- 5) If the answer to 4) is affirmative, then $\mathcal{L}_m(A) \Leftarrow \mathcal{L}_m(P)$. Otherwise not. Halt.

Proof: First observe from the definition of B that a string $t \in \Sigma^*$ is in $\mathcal{L}(B)$ if and only if $\alpha(t, (q_0, \{x_0\})) = (q, \chi)$, where

$$q = \delta(t, q_0) \quad (3.8)$$

and

$$\chi = \{x \in X \mid (\exists i \leq |t|) x = \xi(\text{su}f_i(t), x_0)\}. \quad (3.9)$$

Indeed, (3.8) holds since for each state $(q, \chi') \in V$ the transition $\alpha(\sigma, (q, \chi'))$ is if and only if $\delta(\sigma, q) = q'$. To see that also (3.9) holds, we proceed by induction on the length of strings. For $t = \epsilon$, (3.9) clearly holds with $\chi = \{x_0\}$. Suppose (3.9) holds for an arbitrary string t , so that

$$\chi_t = \{x \in X \mid (\exists i \leq |t|) x = \xi(\text{su}f_i(t), x_0)\}$$

and consider the string $t\sigma$ for an arbitrary $\sigma \in \Sigma$. Employing (3.6) and substituting the above expression for χ_t yields

$$\begin{aligned} \chi_{t\sigma} &= \{x_0\} \cup \{x' \in X \mid (\exists x \in \chi_t) \xi(\sigma, x) = x'\} \\ &= \{x_0\} \cup \{x' \in X \mid (\exists i \leq |t|) \xi(\sigma, \xi(\text{su}f_i(t), x_0)) = x'\} \\ &= \{x_0\} \cup \{x' \in X \mid (\exists i \leq |t\sigma|) \xi(\text{su}f_i(t\sigma), x_0) = x'\} \\ &= \{x' \in X \mid (\exists i \leq |t\sigma|) \xi(\text{su}f_i(t\sigma), x_0) = x'\}. \end{aligned}$$

An immediate consequence of (3.8) and (3.9) is that

$$\mathcal{L}(B) = \mathcal{L}(P) \quad (3.10)$$

and since $V_m = Q_m \times X$

$$\mathcal{L}_m(B) = \mathcal{L}_m(P) \quad (3.11)$$

Suppose now that there exists a state $v = (q, x) \in V$ such that $q \in Q_m$ and $\chi \cap X_m = \phi$. Let $t \in \Sigma^*$ be any string satisfying $\alpha(t, v_0) = v$. By (3.8) it follows that $q = \delta(t, q_0) \in Q_m$, i.e., $t \in \mathcal{L}_m(P)$. By (3.9) it follows that there is no $i \leq |t|$ such that $\xi(\text{su}f_i(t), x_0) \in X_m$. Thus, t has no suffix that belongs to $\mathcal{L}_m(A)$. It follows that $\mathcal{L}_m(P)$ does not converge to $\mathcal{L}_m(A)$. This proves the correctness of step 2).

Suppose next that every state $v = (q, \chi) \in V$ for which $q \in Q_m$ satisfies $\chi \cap X_m \neq \phi$. Then, in view of (3.8) and (3.9), every string $t \in \mathcal{L}_m(P)$ has a suffix that belongs to $\mathcal{L}_m(A)$. Clearly, if $|t| \leq |V - \tilde{T}|$, then the above implies that there exists $i \leq |V - \tilde{T}|$ such that $\text{su}f_i(t) \in \mathcal{L}_m(A)$. We shall show next that when $|V - \tilde{T}|$ is acyclic, there exists $i \leq |V - \tilde{T}|$ such that $\text{su}f_i(t) \in \mathcal{L}_m(A)$ also if $|t| > |V - \tilde{T}|$. If $\tilde{T} = \phi$ this is obvious because then V is acyclic and $|t| \leq |V| = |V - \tilde{T}|$ for all $t \in \mathcal{L}(B)$. When $\tilde{T} \neq \phi$, the acyclicity of $|V - \tilde{T}|$ implies that if $|t| > |V - \tilde{T}|$ then there exists $i < |V - \tilde{T}|$ such that $\alpha(\text{pr}_i(t), v_0) = v = (q, \chi) \in \tilde{T}$. Then, in view of (3.7)

$$\text{su}f_i(t) \in \mathcal{L}_m(P_q) \subset \mathcal{L}_m(A_\chi)$$

or, alternately

$$\text{su}f_i(t) \in \mathcal{L}_m(A_x) \quad (3.12)$$

for some $x \in \chi$. By (3.9) it then follows that there exists $j < i$ such that

$$x = \xi(\text{su}f_j(\text{pr}_i(t)), x_0). \quad (3.13)$$

Upon combining (3.12) and (3.13), we obtain that

$$\begin{aligned} \xi(\text{su}f_i(t), x) &= \\ \xi(\text{su}f_i(t), \xi(\text{su}f_j(\text{pr}_i(t)), x_0)) &= \\ \xi(\text{su}f_j(\text{pr}_i(t)) \cdot \text{su}f_i(t), x_0) &= \\ \xi(\text{su}f_j(t), x_0) \end{aligned}$$

which implies that

$$\text{su}f_j(t) \in \mathcal{L}_m(A). \quad (3.14)$$

Thus, the convergence time of $\mathcal{L}_m(P)$ to $\mathcal{L}_m(A)$ is bounded by $|V - \tilde{T}|$ and hence $\mathcal{L}_m(A) \Leftarrow \mathcal{L}_m(P)$.

To see that the acyclicity of $(V - \tilde{T})$ is necessary for convergence of $\mathcal{L}_m(P)$ to $\mathcal{L}_m(A)$, suppose that there exists a cycle C in $(V - \tilde{T})$. Since B is trim, every state of B is accessible from v_0 . Thus there exists a sequence $\{t_i\}$ of strings, $t_1 < t_2 < t_3 \dots$ such that for each i , $\alpha(t_i, v_0) = v_i = (q_i, \chi_i) \in C$. Suppose now that $\mathcal{L}_m(A) \Leftarrow \mathcal{L}_m(P)$ and let k be the convergence time. Choose a string t_i of the above sequence such that $n := |t_i| > k$. Since no state of C belongs to \tilde{T} , there exists $s \in \mathcal{L}_m(P_{q_i})$ such that $s \notin \mathcal{L}_m(A_x)$ for every $x \in \chi_i$. Now, $t_i s \in \mathcal{L}_m(P)$ and by the convergence assumption of $\mathcal{L}_m(P)$ to $\mathcal{L}_m(A)$, there exists $j \leq k$ such that

$$\text{su}f_j(t_i s) = \text{su}f_j(t_i) \cdot s \in \mathcal{L}_m(A).$$

Letting $x := \xi(\text{su}f_j(t_i), x_0)$, it follows that $s \in \mathcal{L}_m(A_x)$. But by (3.9) $x \in \chi_i$, a contradiction. \square

The complexity of steps 1), 2), and 4) of Algorithm 3.3 is $O(|Q|2^{|X|})$. The test whether $\mathcal{L}_m(P_q) \subset \mathcal{L}_m(A_\chi)$ can be performed by constructing a deterministic recognizer A^χ for $\mathcal{L}_m(A_\chi)$ (an algorithm of complexity $O(2^{|X|})$) and checking whether $\mathcal{L}_m(P_q) \subset \mathcal{L}_m(A^\chi)$ which can be done by a computation of complexity $O(|Q|2^{|X|})$. Thus, the complexity of step 3) as well as that of the complete algorithm is $O((|Q|2^{|X|})^2)$.

Remark 3.1: An alternate approach to verifying the convergence of $\mathcal{L}_m(P)$ to $\mathcal{L}_m(A)$ is based on the reversal of the automata P and A so as to obtain the reflections $\mathcal{L}_m(P)^R$ and $\mathcal{L}_m(A)^R$ of the languages $\mathcal{L}_m(P)$ and $\mathcal{L}_m(A)$. Then $\mathcal{L}_m(A) \Leftarrow \mathcal{L}_m(P)$ if and only if there exists a bounded language L (i.e., a language all of whose strings are of bounded length) such that $\mathcal{L}_m(P)^R \subset \mathcal{L}_m(A)^R \cdot L$. (Here $\mathcal{L}_m(A)^R \cdot L$ denotes the concatenation of the languages $\mathcal{L}_m(A)^R$ and L .) The reversed automata P^R and A^R are nondeterministic in general, and the algorithm for testing the convergence involves computing deterministic equivalents to P^R and A^R . Such an algorithm, whose complexity is $O((2^{|X|} + |Q|)^2)$, has recently been described in [9] and [18]. Algorithm 3.3 above, whose complexity is $O((|Q| \cdot 2^{|X|})^2)$, is superior to the algorithm in [9] and [18] since it is only of polynomial complexity in the size of Q (although it is also of second-degree exponential complexity in the size of X). Thus, Algorithm 3.3 can be regarded as relatively efficient or at least tractable when the specification can be described by a small automaton as opposed to the algorithm of [9] and [18], which is generally impractical.

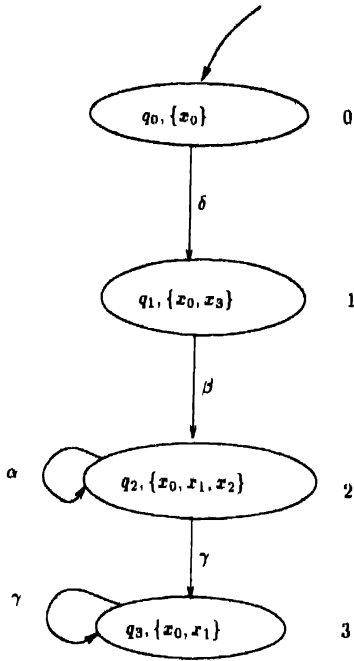


Fig. 4.

Next we illustrate Algorithm 3.3 by employing it to verify the convergence problem of Example 3.2.

Example 3.3 (Verification of Convergence Problem of Example 3.2 — Algorithm 3.3):

Step 1: The automaton B is given in Fig. 4.

To illustrate property (3.9) let us examine state 2). Every string $t \in \Sigma^*$ that satisfies $\alpha(t, v_0) = (q_2, \{x_0, x_1, x_2\})$ is of the form $\delta\beta\alpha^i$ for some $i \geq 0$. If i is even, then the suffixes $\epsilon, \beta\alpha^i, \delta\beta\alpha^i$ satisfy the property $\xi(\epsilon, x_0) = x_0, \xi(\beta\alpha^i, x_0) = x_1$, and $\xi(\delta\beta\alpha^i, x_0) = x_2$. If i is odd, then these suffixes satisfy $\xi(\epsilon, x_0) = x_0, \xi(\beta\alpha^i, x_0) = x_2$, and $\xi(\delta\beta\alpha^i, x_0) = x_1$.

Step 2: Since in the automaton A of Fig. 2 $X_m = X$, we have $\{x\} \cap X_m = \{x\} \cap X = \{x\}$ for each state $x \in X$. Hence $\chi \cap X_m \neq \emptyset$ for each $(q, \chi) \in V$.

Step 3: It is easy to verify that

$$\tilde{T} = \{1, 2, 3\} \neq \emptyset.$$

Step 4: There does not exist a cycle (of B) in $V \cdot \tilde{T} = \{0\}$.

Step 5: $\mathcal{L}_m(A) \Leftarrow \mathcal{L}_m(P)$. \square

IV. LANGUAGE CONVERGENCE IN CONTROLLED DES

In the previous section we examined the problem of verification whether a language L (finitely) converges to a given language M . In the present section we shall deal with the problem of convergence of the language generated by a DES $P = (Q, \Sigma, \delta, q_0, Q_m)$ to a given specification of logical behavior given by a language $E \subset \Sigma^*$. Specifically, we shall examine the problem of existence of a nonblocking supervisor S such that $E \Leftarrow \mathcal{L}_m(S/P)$. Thus we formally define the following.

Controlled Convergence Problem (CCP): Synthesize a nonblocking supervisor S for P such that

$$E \Leftarrow \mathcal{L}_m(S/P). \quad (4.1)$$

We recall from [1] that a language $M, \emptyset \neq M \subset \mathcal{L}_m(P)$ can be realized by a nonblocking supervisor S , i.e., $\mathcal{L}_m(S/P) = M$, if and only if M satisfies the following two conditions:

- M is $\mathcal{L}_m(P)$ -closed, i.e.,

$$M = \bar{M} \cap \mathcal{L}_m(P). \quad (4.2)$$

- M is controllable with respect to $L(P)$, i.e.,

$$\bar{M}\Sigma_u \cap \mathcal{L}(P) \subset \bar{M}. \quad (4.3)$$

We further recall that, for a language N , the class $CL(\mathcal{L}_m(P) \cap N)$ of controllable and $\mathcal{L}_m(P)$ -closed sublanguages of $\mathcal{L}_m(P) \cap N$ has a unique supremal element \hat{N}_p . This implies that a nonblocking supervisor S exists such that $\mathcal{L}_m(S/P) \subset N$ if and only if $\hat{N}_p \neq \emptyset$. In view of the above observations we have the following immediate result.

Proposition 4.1: CCP is solvable if and only if there exists a controllable and $\mathcal{L}_m(P)$ -closed sublanguage $\phi \neq N \subset \mathcal{L}_m(P)$ such that $E \Leftarrow N$.

A special case of interest occurs when the process $P = (Q, \Sigma, \delta, q_0, Q_m)$ satisfies $Q_m = Q$, in which case $\mathcal{L}_m(P) = \mathcal{L}(P)$ and $\mathcal{L}_m(S/P) = \mathcal{L}(S/P)$. Condition (4.2) then becomes that M is prefix closed and, since the class of closed and controllable sublanguages of $\mathcal{L}(P)$ is closed under arbitrary intersections (see e.g., [1]), it contains a unique infimal element. Thus, in view of (2.12) when $Q_m = Q$, the solvability test of CCP simplifies to the following.

Let P^u be the automaton obtained from P upon deleting from it all controlled transitions, i.e., $P^u = (Q, \Sigma, \delta^u, q_0, Q_m)$, where

$$\delta^u(\sigma, q) = \begin{cases} \delta(\sigma, q) & \text{if } \sigma \in \Sigma_u \wedge \delta(\sigma, q) \neq \emptyset \\ \text{undefined} & \text{otherwise.} \end{cases} \quad (4.4)$$

Clearly $\mathcal{L}(P^u)$ is controllable and closed and for any supervisor S

$$\mathcal{L}(P^u) \subset \mathcal{L}(S/P).$$

Proposition 4.1 can now be replaced (for the case $Q_m = Q$) by the following proposition.

Proposition 4.2: Let $P = (Q, \Sigma, \delta, q_0, Q_m)$ with $Q_m = Q$ and let $E \subset \Sigma^*$ be given. Then CCP is solvable if and only if $E \Leftarrow \mathcal{L}(P^u)$.

The above proposition yields a simple procedure for solving CCP in the special case when $Q_m = Q$. First, one constructs P^u and checks whether $E \Leftarrow \mathcal{L}(P^u)$. This verification can be done with Algorithm 3.2 (which is of polynomial complexity) when E is suffix closed, and with Algorithm 3.3 otherwise. When $E \Leftarrow \mathcal{L}(P^u)$, then the most restrictive supervisor, i.e., the supervisor that disables all controlled events, is a solution of CCP because this supervisor satisfies

$$E \Leftarrow \mathcal{L}(S/P) = \mathcal{L}(P^u).$$

The above solution is, of course, not a very efficient one and in general better solutions might be sought. We shall return to this issue later.

Remark 4.1: In [9] and [18], a problem similar to CCP was discussed. A language L was called l -stabilizable w.r.t. E if there exists a sublanguage $H \subset L$, controllable w.r.t. L , such that $E \Leftarrow H$. The authors argued there that for a system $P = (Q, \Sigma, \delta, q_0, Q_m)$, $\mathcal{L}_m(P)$ is l -stabilizable w.r.t. a given language E if and only if $E \Leftarrow \mathcal{L}_m(P^u)$. The justification for this conclusion was the claim that $\mathcal{L}_m(P^u)$ is the infimal controllable sublanguage of $\mathcal{L}_m(P)$. When $Q_m = Q$, this claim is indeed true. It is false, however, in the general case when $Q_m \neq Q$, since in that case $\mathcal{L}_m(P^u)$ is not necessarily a controllable language. To solve the l -stabilizability problem as well as CCP in the general case, a different approach must be taken as discussed in detail below.

We turn next to the problem of solving CCP in the general case when $Q_m \neq Q$ and E is a general regular language. To this end we shall make use of two known algorithms that are briefly discussed below.

The first algorithm computes the supremal controllable and $\mathcal{L}_m(P)$ -closed sublanguages of a given language. Specifically, let $P = (Q, \Sigma, \delta, q_0, Q_m)$ be a DES and $M \subset \mathcal{L}_m(P)$ be a given regular language. Let $A = (X, \Sigma, \xi, x_0, X_m)$ be a deterministic recognizer for M ($\mathcal{L}_m(A) = M$). The construction of a deterministic recognizer for the supremal controllable (with respect to $\mathcal{L}(P)$) and $\mathcal{L}_m(P)$ -closed sublanguage of M is accomplished by an algorithm of complexity of $O(|Q||X|)$ as follows. First construct the synchronous composition $C' = P||A = (Q \times X, \Sigma, \alpha, (q_0, x_0), Q_m \times X_m)$ as described in Section II. Then $\mathcal{L}_m(C') = \mathcal{L}_m(P) \cap \mathcal{L}_m(A) = \mathcal{L}_m(P) \cap M = M$. The supremal $\mathcal{L}_m(P)$ -closed sublanguage of M is then given by $\mathcal{L}_m(C')$ (see [14, Appendix B]), where C' is given by the restriction

$$C' = C'|_{Q_m \times (X - X_m)}.$$

The complexity of constructing C' is $O(|Q||X|)$. By [2, Proposition, 6.1], the supremal controllable sublanguage of $\mathcal{L}_m(C')$ equals the supremal controllable and $\mathcal{L}_m(P)$ -closed sublanguage of M . A deterministic recognizer for the supremal controllable sublanguage of $\mathcal{L}_m(C')$ can be constructed by the algorithm of [11] (see also [12]), whose complexity is $O(|Q|^2 \cdot |X|^2)$. An algorithm of complexity $O(|Q||X|)$ has recently been given in [13]. Thus, the construction of the supremal controllable and $\mathcal{L}_m(P)$ -closed sublanguage of M can be accomplished with overall complexity $O(|Q||X|)$.

The second algorithm that we shall employ below appeared in [4] (see also [5], [9], [18]) and is briefly described next. Let $P = (Q, \Sigma, \delta, \cdot, Q_m)$ be a DES with unspecified initial state. Fix a subset $\hat{Q} \subset Q$ (which we shall regard as the set of legal states). The set \hat{Q} is called a strong attractor for a state $q \in Q$ if, when initialized at q , the system P always reaches a state in \hat{Q} after a bounded number of transitions. Formally, \hat{Q} is a strong attractor for $q \in Q$ if $\mathcal{A}(P|\hat{Q}, q)$ is acyclic and no state of $\mathcal{A}(P|\hat{Q}, q)$ is a deadlock state of P (i.e., a state at which no transition is defined). A set \hat{Q} is called a weak attractor for $q \in Q$ if there exists a supervisor S such that \hat{Q} is a strong attractor for q in (S/P) . Such a supervisor, when it exists, can be chosen to be static, i.e., for $s, t \in \mathcal{L}(P)$, $S(t) = S(s)$ whenever $\delta(s, q_0) = \delta(t, q_0) (= q)$. Note also that a static supervisor can be written as a map

$S : X \rightarrow 2^\Sigma$. The set of all states $q \in Q$ for which \hat{Q} is a weak attractor is called the region of weak attraction and is denoted by $\Omega_P(\hat{Q})$. If $\hat{Q} = \phi$, then $\Omega_P(\hat{Q}) := \phi$. In [4] and [5] algorithms are given for computation of $\Omega_P(\hat{Q})$ with complexity $O(|Q|^2)$. An algorithm of complexity $O(|Q|)$ has recently been presented in [9] and [18].

We return to our main discussion. To test the solvability of CCP, we first construct the automaton $B = (V, \Sigma, \alpha, v_0, V_m)$ as defined in Step 1) of Algorithm 3.3. In view of the fact that $\mathcal{L}(B) = \mathcal{L}(P)$ and $\mathcal{L}_m(B) = \mathcal{L}_m(P)$, [(3.10), (3.11)], it is clear that B can be regarded as a new model of our process. Thus, CCP is solvable with respect to P if and only if it is solvable with respect to B .

The automaton B offers an obvious advantage over P for testing the solvability of CCP in that its structure provides a simple way of identifying all strings of $\mathcal{L}_m(B)$ that do not have a suffix in E . Indeed, let $t \in \mathcal{L}_m(B)$ and $v = (q, \chi) \in V_m$ satisfy $\alpha(t, (q_0, \{x_0\})) = v$. Then, since $V_m = Q_m \times 2^X$, it follows that $q \in Q_m$ and, see (3.9), t has a suffix in E if and only if $\chi \cap X_m \neq \phi$. Let F be the set defined by

$$F = \{(q, \chi) \in V_m | q \in Q_m \text{ and } \chi \cap X_m \neq \phi\}. \quad (4.5)$$

Define B' as the automaton

$$B' = (V, \Sigma, \alpha, v_0, F)$$

which is obtained from B by restricting its set of final states to F . Suppose now that S is a solution to CCP with respect to B , i.e., $E \Leftarrow \mathcal{L}_m(S/B)$, and let $t \in \mathcal{L}_m(B) - \mathcal{L}_m(B')$. Since t has no suffix in E , it follows that $t \notin \mathcal{L}_m(S/B)$ and it follows further that $\mathcal{L}_m(S/B) \subset \mathcal{L}_m(B')$. Since $\mathcal{L}_m(S/B)$ is controllable with respect to $\mathcal{L}(B)$ and is $\mathcal{L}_m(B)$ -closed, we conclude that $\mathcal{L}_m(S/B) \subset \sup CL(\mathcal{L}_m(B'))$, where $\sup CL(\mathcal{L}_m(B'))$ is the supremal controllable (w.r.t. $\mathcal{L}(B)$) and $\mathcal{L}_m(B)$ -closed sublanguage of $\mathcal{L}_m(B')$. Define now the automaton

$$\hat{B} = (\hat{V}, \Sigma, \hat{\alpha}, \hat{v}_0, \hat{V}_m) \quad (4.6)$$

as a recognizer for $\sup CL(\mathcal{L}_m(B'))$. Such an automaton can be constructed by an algorithm of complexity $O(|Q| \cdot 2^{|X|})$ as was described earlier. From the algorithm it is also clear that \hat{B} is a subautomaton of B . Now, the inclusion $\mathcal{L}_m(S/B) \subset \mathcal{L}_m(\hat{B}) \subset \mathcal{L}_m(B)$ implies that $\mathcal{L}_m(S/B) = \mathcal{L}_m(S/\hat{B})$. Hence, CCP is solvable with respect to B only if it is solvable with respect to \hat{B} . To see the reverse implication, i.e., that solvability of CCP with respect to \hat{B} implies the solvability of CCP with respect to B , observe that $\mathcal{L}_m(\hat{B})$ is a controllable (w.r.t. $\mathcal{L}(B)$) and $\mathcal{L}_m(B)$ -closed sublanguage of $\mathcal{L}_m(B)$. Thus, there exists a supervisor \hat{S} such that $\mathcal{L}_m(\hat{S}/B) = \mathcal{L}_m(\hat{B})$. If there exists a supervisor S such that $E \Leftarrow \mathcal{L}_m(S/\hat{B})$, then $E \Leftarrow \mathcal{L}_m(S/(\hat{S}/B)) = \mathcal{L}_m(S \times \hat{S}/B)$, where $S \times \hat{S}$ is the supervisor that consists of the conjunction of S and \hat{S} (see e.g., [2]). We have just proved the following.

Lemma 4.1: For a DES P and a regular language E , let \hat{B} be the automaton defined by (4.6). Then CCP is solvable with respect to P if and only if CCP is solvable with respect to \hat{B} .

The following corollary is immediate.

Corollary 4.1: If $\mathcal{L}_m(\hat{B}) = \emptyset$ then CCP is unsolvable.

In Theorem 3.2 (see Algorithm 3.3) we have shown that the convergence $\mathcal{L}_m(A) \Leftarrow \mathcal{L}_m(P)$ is equivalent to the acyclicity of $V - \hat{T}$ where \hat{T} [see (3.7)] consists of all states $v = (q, \chi)$ such that $\mathcal{L}_m(P_q) \subset \mathcal{L}_m(A_\chi)$. Similarly, we construct now a subset \hat{T} of states of \hat{B} such that for each $v = (q, \chi) \in \hat{T}$ there exists a supervisor S with the property that $\mathcal{L}_m(S/P_q) \subset \mathcal{L}_m(A_\chi)$. We will then show that the solvability of CCP is equivalent to the possibility of preventing the system \hat{B} from executing cycles in $\hat{V} - \hat{T}$.

For a state $v = (q, \chi)$ of \hat{B} it follows from (4.2) and (4.3) that there exists a nonblocking supervisor S such that $\mathcal{L}_m(S/P_q) \subset \mathcal{L}_m(A_\chi)$ if and only if the class of controllable (with respect to $\mathcal{L}(P_q)$) and $\mathcal{L}_m(P_q)$ -closed sublanguages of $\mathcal{L}_m(P_q) \cap \mathcal{L}_m(A_\chi)$ does not consist of the empty language. To check this condition we proceed as follows. First note that A_χ is not a deterministic automaton since it has in general more than one initial state. Hence we begin by constructing a deterministic recognizer $A^\chi = (W, \Sigma, \beta, w_0, W_m)$ for $\mathcal{L}_m(A_\chi)$, i.e., $\mathcal{L}_m(A^\chi) = \mathcal{L}_m(A_\chi)$. Next, by the algorithm described earlier we construct a deterministic automaton D^ν such that

$$\mathcal{L}_m(D^\nu) = \sup CL(\mathcal{L}_m(P_q) \cap \mathcal{L}_m(A^\chi)). \quad (4.7)$$

The required supervisor S such that $\mathcal{L}_m(S/P_q) \subset \mathcal{L}_m(A_\chi)$ exists if and only if $\mathcal{L}_m(D^\nu) \neq \emptyset$. The construction of D^ν requires $O(|Q| \cdot 2^{|X|})$ computations since the number of states of A^χ is bounded by $2^{|X|}$ and the number of states of P_q is $|Q|$.

Now, let \hat{T} be the set of states defined by

$$\hat{T} = \{v \in \hat{V} \mid \mathcal{L}_m(D^\nu) \neq \emptyset\}. \quad (4.8)$$

Since the number of states of \hat{B} is bounded by $|Q| \cdot 2^{|X|}$ (the number of states of B) the construction of \hat{T} requires $O((|Q| \cdot 2^{|X|})^2)$ computations.

The final step in our computation is the employment of the algorithm of [9] and [18] to compute the set $\Omega_B(\hat{T})$. The complexity of this algorithm is $O(|Q| \cdot 2^{|X|})$.

We now have the following necessary and sufficient condition for the solvability of CCP.

Theorem 4.1: Let P be a DES, and let $E \subset \Sigma^*$ be a regular language. Let \hat{B} and \hat{T} be the automaton and the set of states as defined in (4.6) and (4.8), respectively. Then CCP is solvable if and only if $v_0 \in \Omega_B(\hat{T})$.

Proof: Only If. By Lemma 4.1 CCP is solvable with respect to P if and only if it is solvable with respect to \hat{B} . Suppose that CCP is solvable so that there exists a nonblocking supervisor S such that $E \Leftarrow \mathcal{L}_m(S/\hat{B})$. We shall show that the assumption that $v_0 \notin \Omega_B(\hat{T})$ leads to a contradiction. First we shall show that, under S , there are no deadlock states in $\hat{V} - \hat{T}$. Indeed, let $\hat{v} = (q, \chi) \in \hat{V} - \hat{T}$ and $t \in \mathcal{L}(S/\hat{B})$ be such that $\hat{\alpha}(t, \hat{v}_0) = \hat{v}$. Suppose that \hat{v} is a deadlock state, i.e., $t_s \in$

$\mathcal{L}(S/\hat{B}) \Rightarrow s = \epsilon$. If $q \in Q - Q_m$, then $(q, \chi) \notin \hat{V}_m$, implying that $t \notin \mathcal{L}_m(S/\hat{B})$. Thus $\mathcal{L}(S/\hat{B}) \neq \mathcal{L}_m(S/\hat{B})$, contradicting the nonblocking property of S . If $q \in Q_m$ then $\epsilon \in \mathcal{L}_m(P_q)$. By definition of \hat{B} , $\chi \cap X_m \neq \emptyset$ so that $\epsilon \in \mathcal{L}_m(A_\chi)$, implying that $\epsilon \in \mathcal{L}_m(P_q) \cap \mathcal{L}_m(A_\chi)$. Since the language $\{\epsilon\}$ is trivially $\mathcal{L}_m(P_q)$ -closed and controllable with respect to $\mathcal{L}(P_q)$, it follows that $\{\epsilon\} \in CL(\mathcal{L}_m(P_q) \cap \mathcal{L}_m(A_\chi))$, so that $\mathcal{L}_m(D^\nu) \neq \emptyset$. This implies that $\hat{v} \in \hat{T}$, a contradiction.

Since (S/\hat{B}) has no deadlock states in $\hat{V} - \hat{T}$, the assumption that $\hat{v}_0 \notin \Omega_B(\hat{T})$ implies that (S/B) has a cycle in $\hat{V} - \hat{T}$. With the employment of the same arguments as in the proof of Theorem 3.2, it is not difficult to show that the existence of a cycle in $\hat{V} - \hat{T}$ contradicts the assumption that $E \Leftarrow \mathcal{L}_m(S/\hat{B})$.

If. The proof of the "if" part of the theorem consists of the following algorithm for construction of the supervisor S that satisfies the condition that $E \Leftarrow \mathcal{L}_m(S/\hat{B})$ whenever $\hat{v}_0 \in \Omega_B(\hat{T})$.

Algorithm 4.1:

Step 1) Define the supervisor $S_1 : V \rightarrow 2^{\Sigma^*}$ as a static supervisor such that T is a strong attractor for $\Omega_B(\hat{T})$ with respect to S_1/\hat{B} . The existence of this supervisor follows from the definition of $\Omega_B(\hat{T})$.

Step 2) For each $\hat{v} \in \hat{V}$, let D^ν be the automaton defined earlier (see (4.7)).

The required supervisor is given by $S : L(\hat{B}) \rightarrow 2^{\Sigma^*}$ as found in (4.9) at the bottom of the page.

The supervisor S is obviously well defined because for each $t \in \mathcal{L}(\hat{B})$ it defines a unique set of controlled events to be disabled. Next, note that under the supervision of S , the system B reaches a state $v \in T$ within a number of transitions bounded by $|\hat{V} - \hat{T}|$ because for states in $\Omega_B(\hat{T}) - \hat{T}$ S acts as S_1 . When a state $\hat{v} = (q, \chi) \in T$ is reached, then it is guaranteed by S that all marked strings generated by \hat{B} (starting at \hat{v}) belong to

$$\mathcal{L}_m(D^\nu) = \sup CL(\mathcal{L}_m(P_q) \cap \mathcal{L}_m(A^\chi)) \subset \mathcal{L}_m(A_\chi).$$

It thus follows that for $t \in \mathcal{L}_m(S/\hat{B})$, $|t| \geq |\hat{V} - \hat{T}|$, there exists $i \leq |\hat{V} - \hat{T}|$ such that $\alpha(pr_i(t), \hat{v}_0) = \hat{v} = (q, \chi) \in \hat{T}$ and $\text{suf}_i(t) \in \mathcal{L}_m(A_\chi)$. By (3.12)–(3.14) it follows that there exists $j \leq i$ such that $\text{suf}_j(t) \in E$. If $|t| < |\hat{V} - \hat{T}|$, the existence of i ($\leq |\hat{V} - \hat{T}|$) such that $\text{suf}_i(t) \in E$ follows from the definition of \hat{B} . This concludes the proof. \square

The complexities of computing the automaton \hat{B} and the set \hat{T} are $O(|Q| \cdot 2^{|X|})$ and $O((|Q| \cdot 2^{|X|})^2)$, respectively. The computation of $\Omega_B(\hat{T})$ requires $O(|Q| \cdot 2^{|X|})$ computations. Thus, the complexity of verifying the solvability of CCP (i.e., checking the condition $\hat{v}_0 \in \Omega_B(\hat{T})$) is $O((|Q| \cdot 2^{|X|})^2)$. In case CCP is solvable, the supervisor defined by (4.9) is a solution of CCP.

$$\begin{aligned} S_1(\hat{v}) & \quad \text{where } \hat{v} = \hat{\alpha}(t, \hat{v}_0) & \quad \text{if } \{\forall_i\} \hat{\alpha}(pr_i(t), \hat{v}_0) \notin \hat{T} \\ S(t) = & \quad \{\sigma \in \Sigma_c \mid \text{suf}_{i_0}(t)\sigma \notin \mathcal{L}(D^\nu)\}, & \quad \text{if there exists } i \text{ such that} \\ & \quad \text{where } i_0 \text{ is the least integer } i & \quad \hat{\alpha}(pr_i(t), \hat{v}_0) \in \hat{T} \\ & \quad \text{that satisfies } \hat{\alpha}(pr_i(t), \hat{v}_0) \in \hat{T} & \end{aligned} \quad (4.9)$$

When E is suffix closed, the solvability of CCP can be verified by an algorithm of linear complexity (i.e., of complexity $O(|Q||X|)$) as described below. The correctness of the following algorithm follows from Theorems 3.1 and 4.1.

Theorem 4.2: Let $P = (Q, \Sigma, \delta, q_0, Q_m)$ be a given DES, and let E be a suffix-closed regular language. Let $A = (X, \Sigma, \xi, x_0, X_m)$ be a deterministic automaton such that $\mathcal{L}_m(A) = E$. Then the following algorithm verifies the solvability of CCP.

Algorithm 4.2:

- Step 1) For each $q \in Q$ construct a deterministic automaton D^q such that $\mathcal{L}_m(D^q)$ is the supremal controllable (w.r.t. $\mathcal{L}(P_q)$) and $\mathcal{L}_m(P_q)$ -closed sublanguage of $\mathcal{L}_m(P_q) \cap \mathcal{L}_m(A)$.
- Step 2) Compute the set Y of states defined by

$$Y := \{q \in Q \mid \mathcal{L}_m(D^q) \neq \emptyset\}.$$

- Step 3) By the algorithm of [9], [18] compute the set $\Omega_P(Y)$.

- Step 4) CCP is solvable if and only if $q_0 \in \Omega_P(Y)$. \square

The set Y can be constructed with the algorithm of [13] using the same method as was described in Section III for the computation of the set T . The complexity of constructing the set Y as well as the overall complexity of Algorithm 4.2 is $O(|Q||X|)$.

Algorithms 4.1 and 4.2, which constitute tests for solvability of CCP, also provide a construction of nonblocking supervisors that guarantee convergence of the closed-loop language to the specification E whenever CCP is solvable. The synthesis is, however, neither optimal nor unique since, in general, there does not exist a least restrictive supervisor that guarantees convergence. Specifically, there does not, in general, exist a supervisor S such that $E \Leftarrow \mathcal{L}_m(S/P)$ and such that if S is any supervisor satisfying $E \Leftarrow \mathcal{L}_m(S/P)$, then $\mathcal{L}_m(S/P) \subset \mathcal{L}_m(\hat{S}/P)$. To see this, consider the following simple example. Let $\mathcal{L}_m(P) = \alpha^*\beta$ and let $E = \beta$. Assume further that $\Sigma_c = \{\alpha\}$ and $\Sigma_u = \{\beta\}$. The solvability of CCP is obvious, and for any $k \geq 0$ there is a supervisor S_k such that $\mathcal{L}_m(S_k/P) = \alpha^k\beta$ (which clearly satisfies $\beta \Leftarrow \alpha^k\beta$). But for $k_1 \leq k_2$, $\alpha^{k_1}\beta \subset \alpha^{k_2}\beta$ and k can be arbitrarily large!

The issue of optimal supervisors (in contexts other than "minimally restrictive") is discussed in [17].

V. ASYMPTOTIC BEHAVIOR OF DES

In nonterminating systems, i.e., systems that operate indefinitely, one can distinguish between their transient and permanent (or asymptotic) behaviors. Intuitively, the asymptotic behavior of a system consists of all strings that the system can execute after having already performed an arbitrarily large number of transitions.

For a language $L \subset \Sigma^*$ we define its asymptotic suffix, denoted L_∞ as

$$L_\infty := \{s \in \Sigma^* \mid \forall l \geq 0, \exists t_l \in \Sigma^*, |t_l| \geq l \wedge t_l s \in L\}. \quad (5.1)$$

Clearly, L_∞ is an empty language whenever L is a bounded language (see Remark 3.1). It is also easy to see that L_∞ is suffix closed for every language $L \subset \Sigma^*$.

Let $P = (Q, \Sigma, \delta, q_0, Q_m)$ be a DES. The asymptotic behavior of P is defined as $\mathcal{L}_m(P)_\infty$. A recognizer for $\mathcal{L}_m(P)_\infty$ can be constructed as follows. Let $cy(P)$ be the set of all states of P that are accessible from q_0 and belong to a cycle of P , that is

$$cy(P) := \{q \in Q \mid (\exists s, t \in \Sigma^*) q = \delta(t, q_0), s \neq \epsilon \wedge \delta(s, q) = q\}. \quad (5.2)$$

Let $Q_\infty := \mathcal{A}(P, cy(P))$ and let $Q_{m\infty} := Q_m \cap Q_\infty$. Define

$$P_\infty := (Q_\infty, \Sigma, \delta, Q_\infty, Q_{m\infty}). \quad (5.3)$$

The following proposition states that P_∞ is a recognizer for $\mathcal{L}_m(P)_\infty$.

Proposition 5.1: $\mathcal{L}_m(P)_\infty = \mathcal{L}_m(P_\infty)$.

Proof: $\mathcal{L}_m(P)_\infty \subset \mathcal{L}_m(P_\infty)$. Let $t \in \mathcal{L}_m(P)_\infty$ be any string. By (5.1) there exists a string $w, |w| > |Q|$, such that $wt \in \mathcal{L}_m(P)$. Let $q_0, q_1, \dots, q_l, l = |w|$, be the path associated with w . Since $l > |Q|$, there exists a state $q \in Q$ that occurs along this path at least twice, i.e., there exist $i_1, i_2, 0 \leq i_1 < i_2 \leq l$ such that $q = q_{i_1} = q_{i_2}$. Thus, $q \in cy(P)$ and $q' := \delta(w, q_0) \in \mathcal{A}(P, cy(P)) = Q_\infty$. Since $wt \in \mathcal{L}_m(P)$ and P is deterministic, it follows that

$$t \in \mathcal{L}_m(Q_\infty, \Sigma, \delta, q', Q_{m\infty}) \subset \mathcal{L}_m(Q_\infty, \Sigma, \delta, Q_\infty, Q_{m\infty}) = \mathcal{L}_m(P_\infty).$$

$\mathcal{L}_m(P)_\infty \supset \mathcal{L}_m(P_\infty)$. Let $w \in \mathcal{L}_m(P_\infty)$. Then $w \in \mathcal{L}_m(Q_\infty, \Sigma, \delta, q, Q_{m\infty})$ for some $q \in Q_\infty$. Thus see the equation at the bottom of the page.

By (5.1) this implies that $w \in \mathcal{L}_m(P)_\infty$ concluding the proof. \square

The construction of the set $cy(P)$ can be accomplished by the well-known algorithm for computing strongly connected components of P (see [7, p. 64]) whose complexity (under our assumption that $|\Sigma| = \mathcal{O}(1)$) is $O(|Q|)$. The construction of $\mathcal{A}(P, cy(P))$ requires $O(|Q|)$ computations. Thus, the complexity of the construction of P_∞ is linear in the number of states of P .

Our next goal will be to show that for a given system P , the convergence of $\mathcal{L}_m(P)$ to a given language E can be determined by testing the asymptotic behavior of P . The following proposition gives a necessary condition for the convergence $E \Leftarrow \mathcal{L}_m(P)$ in terms of the asymptotic behavior of P .

Proposition 5.2: Let P be a given DES, and let $E \subset \Sigma^*$ be a given language. If $E \Leftarrow \mathcal{L}_m(P)$, then $\mathcal{L}_m(P)_\infty \subset \text{suf}(E)$.

$$\begin{aligned} & \exists u \in \Sigma^*, \exists q' \in cy(P), \delta(u, q') = q && \text{(definition of } Q_\infty) \\ \Rightarrow & \exists (t, s \in \Sigma^*) q' = \delta(t, q_0), s \neq \epsilon \text{ and } q' = \delta(s, q') && \text{(follows by (5.2))} \\ \Rightarrow & ts^*uw \in \mathcal{L}_m(P). \end{aligned}$$

Proof: Assume that $E \Leftarrow \mathcal{L}_m(P)$, let the convergence time be k and let $s \in \mathcal{L}_m(P)_\infty$ be any string. By (5.1), for each $i \geq 0$ there exists $t_i \in \Sigma^*$, $|t_i| \geq i$, such that $t_i s \in \mathcal{L}_m(P)$. Choose $t_i > k$. Since there exists $j \leq k$ such that $\text{suf}_j(t_i s) \in E$, it follows that

$$s = \text{suf}_{|t_i|}(t_i s) = \text{suf}_{|t_i|-j}(\text{suf}_j(t_i s)) \in \text{suf}(E)$$

concluding the proof. \square

The following lemma states that every regular language $L \subset \Sigma^*$ converges to L_∞ .

Lemma 5.1: For a DES $P = (Q, \Sigma, \delta, q_0, Q_m)$, $\mathcal{L}_m(P)_\infty \Leftarrow \mathcal{L}_m(P)$.

Proof: Since the language $\mathcal{L}_m(P)_\infty$ is suffix closed it follows that $\epsilon \in \mathcal{L}_m(P)_\infty$. Thus for each $t \in \mathcal{L}_m(P)$, $\text{suf}_{|t|}(t) = \epsilon \in \mathcal{L}_m(P)_\infty$ implying that if $|t| < |Q|$ there exists $i \leq |Q|$ such that $\text{suf}_i(t) \in \mathcal{L}_m(P)_\infty$.

Let $t \in \mathcal{L}_m(P)$, $|t| > |Q|$. Since the number of states of P is $|Q|$, there exists $i, i \leq |Q|$, such that $\delta(\text{pr}_i(t), q_0) = q$ where $q \in \text{cy}(P) \subset Q_\infty$. Since by Proposition 5.1 $\mathcal{L}_m(P)_\infty = \mathcal{L}_m(P_\infty)$, it then follows that $\text{suf}_i(t) \in \mathcal{L}_m(P_\infty) = \mathcal{L}_m(P)_\infty$. Hence the convergence time of $\mathcal{L}_m(P)$ to $\mathcal{L}_m(P)_\infty$ is bounded by $|Q|$, concluding the proof. \square

It should be noted that if L is not a regular language, then the convergence $L_\infty \Leftarrow L$ does not necessarily hold. Consider, for example, the language $\{\alpha^k \beta^k \gamma^* \mid \forall k \geq 0\}$.

Proposition 5.2 and Lemma 5.1 yield the following.

Corollary 5.1: Let P be a DES. Then $\mathcal{L}_m(P)_\infty$ is the infimal (in the sense of language inclusion) suffix-closed language that $\mathcal{L}_m(P)$ converges to, i.e.,

$$\mathcal{L}_m(P)_\infty = \cap \{M \subset \Sigma^* \mid \text{suf}(M) = M \text{ and } M \Leftarrow \mathcal{L}_m(P)\}.$$

A sufficient condition for the convergence $E \Leftarrow \mathcal{L}_m(P)$ is given by the following proposition.

Proposition 5.3: For a DES P and a language $E \subset \Sigma^*$, if $\mathcal{L}_m(P)_\infty \subset E$, then $E \Leftarrow \mathcal{L}_m(P)$.

Proof: By Lemma 5.1, $\mathcal{L}_m(P)_\infty \Leftarrow \mathcal{L}_m(P)$. Thus, if $\mathcal{L}_m(P)_\infty \subset E$, it follows directly by (2.12) that also $E \Leftarrow \mathcal{L}_m(P)$. \square

By combining Propositions 5.2 and 5.3 for the special case of suffix-closed languages we obtain the following necessary and sufficient condition for $E \Leftarrow \mathcal{L}_m(P)$.

Theorem 5.1: Let P be a DES and let $E \subset \Sigma^*$ be a suffix-closed language. Then $E \Leftarrow \mathcal{L}_m(P)$ if and only if

$$\mathcal{L}_m(P)_\infty \subset E. \quad (5.4)$$

Next we turn to the problem of supervisory control. In the previous section we considered the problem of language convergence, that is, the problem of synthesizing a supervisor that guarantees convergence of the supervised language to a specified legal language. We shall now be interested in the less restrictive control problem wherein only the asymptotic behavior of the supervised process is required to lie in a specified legal language. Thus, we shall consider the following.

Asymptotic Control Problem (ACP): Let P be a given DES and let $E \subset \Sigma^*$ be a given language. Synthesize a nonblocking supervisor S such that

$$\mathcal{L}_m(S/P)_\infty \subset E. \quad (5.5)$$

In the case when E is a suffix-closed language there is an obvious equivalence between ACP and CCP. This equivalence is stated in the following reinterpretation of Theorem 5.1.

Proposition 5.4: Let E be suffix closed. Then ACP is solvable if and only if CCP is solvable.

Let us now examine the solvability of ACP in the case when E is not suffix closed. Since L_∞ is suffix closed for any language L , ACP is solvable (i.e., there exists a nonblocking supervisor that satisfies (5.5)) if and only if there exists a suffix-closed sublanguage $M \subset E$ and a nonblocking supervisor S such that

$$\mathcal{L}_m(S/P)_\infty \subset M. \quad (5.6)$$

By Theorem 5.1, a supervisor S satisfies (5.6) if and only if

$$M \Leftarrow \mathcal{L}_m(S/P). \quad (5.7)$$

Thus, to determine the solvability of ACP, it is sufficient to check whether CCP is solvable with respect to the supremal suffix-closed sublanguage of E . The existence of the supremal suffix-closed sublanguage of E follows from the closeness of the set of suffix-closed languages under arbitrary unions.

Proposition 5.5: For a language E , ACP is solvable if and only if CCP is solvable with respect to the supremal suffix-closed sublanguage of E .

We conclude this section with an algorithm for computing the supremal suffix-closed sublanguage of a given regular language E .

Algorithm 5.1:

Input) A deterministic automaton

$$A = (X, \Sigma, \xi, x_0, X_m)$$

such that (without loss of generality) $\xi(\sigma, x) \neq x_0$ for all $\sigma \in \Sigma$, $x \in X$.

Output) A deterministic recognizer for the supremal suffix-closed sublanguage of $\mathcal{L}_m(A)$.

Construct a deterministic automaton

$$C = (U, \Sigma, \beta, u_0, U_m)$$

where $U = 2^X$, $u_0 = \{x_0\}$, $U_m = \{u \in U \mid (u - \{x_0\}) \subset X_m\}$ and $\beta: \Sigma \times U \rightarrow U$ is defined by

$$\beta(\sigma, u) = \begin{cases} \{x_0\} \cup \{\xi(\sigma, x) \mid x \in u\} & \text{if } \xi(\sigma, x)! \text{ for all } x \in u \\ \text{undefined} & \text{otherwise.} \end{cases} \quad (5.8)$$

Proposition 5.6: $\mathcal{L}_m(C)$ is the supremal suffix-closed sublanguage of $\mathcal{L}_m(A)$.

Proof: First observe that for any two states $u, w \in U$ such that $u \subset w$ (as elements of 2^X), $\beta(\sigma, w)!$ only if $\beta(\sigma, u)!$, and if $\beta(\sigma, w)!$ then $\beta(\sigma, u) \subset \beta(\sigma, w)$. Moreover, if $\beta(\sigma, u)!$ and $x \in u$, then $\xi(\sigma, x)!$

Let $s = \sigma_1 \cdots \sigma_n \in \mathcal{L}_m(C)$ be any string, and let u_0, \dots, u_n be the associated path in C , i.e., $u_i = \beta(\sigma_i, u_{i-1})$, $i = 1, \dots, n$ and $u_n \in U_m$. Then since $u_0 = \{x_0\}$, it follows from the above observation that there exists a path x_0, \dots, x_n in A such that $x_i = \xi(\sigma_i, x_{i-1})$ and

$x_i \in u_i - \{x_0\}$ for all $i = 1, \dots, n$. From the definition of U_m , $x_n \in u_n - \{x_0\} \subset X_m$, so that $s \in \mathcal{L}_m(A)$, implying that $\mathcal{L}_m(C) \subset \mathcal{L}_m(A)$.

To see that $\mathcal{L}_m(C)$ is suffix closed, consider $\text{suf}_j(s) = \sigma_{j+1} \dots \sigma_n$ for some $0 \leq j \leq n$. We must show that $\text{suf}_j(s) \in \mathcal{L}_m(C)$. Note that $u_0 \subset u_j$, whence since $u_{j+1} = \beta(\sigma_{j+1}, u_j)$ is defined, so is $\hat{u}_1 := \beta(\sigma_{j+1}, u_0)$ and $\hat{u}_1 \subset u_{j+1}$. Proceeding inductively, we obtain a sequence of states $u_0, \hat{u}_1, \dots, \hat{u}_{n-j}$ with $\hat{u}_1 = \beta(\sigma_{j+1}, u_0)$, $\hat{u}_{i+1} = \beta(\sigma_{i+j+1}, \hat{u}_i)$ and $\hat{u}_i \subset u_{i+j}$. Since $\hat{u}_{n-j} \subset u_n \in U_m$, we conclude that $\hat{u}_{n-j} \in U_m$, whence $\text{suf}_j(s) \in \mathcal{L}_m(C)$ as claimed.

Finally, to see that $\mathcal{L}_m(C)$ is the supremal suffix-closed sublanguage of $\mathcal{L}_m(A)$, we must show that if $s = \sigma_1 \dots \sigma_n \in \mathcal{L}_m(A)$ is a string such that $\text{suf}_j(s) \in \mathcal{L}_m(A)$ for all $j = 0, \dots, n$, then $s \in \mathcal{L}_m(C)$. Suppose that $s \notin \mathcal{L}_m(C)$. Let $l, 1 \leq l \leq n$, be the smallest integer such that $u_j = \beta(\sigma_j, u_{j-1})$ is defined for all $j = 1, \dots, l-1$ but $\beta(\sigma_l, u_{l-1})$ is undefined. This implies [see (5.8)] that there exists $k, 0 \leq k \leq l-1$, and a sequence of states $x_0, x_1, \dots, x_{l-k-1}$ such that $x_j = \xi(\sigma_{k+j}, x_{j-1})$ for all $j = 1, \dots, l-k-1$ and $\xi(\sigma_l, x_{l-k-1})$ is undefined. But then $\text{suf}_k(s) \notin \mathcal{L}(A)$, a contradiction. \square

REFERENCES

- [1] P. J. Ramadge and W. M. Wonham, "Supervisory control of a class of discrete-event processes," *SIAM J. Control Optim.*, vol. 25, no. 1, pp. 206-230, Jan. 1987.
- [2] ———, "Modular supervisory control of discrete-event systems," *Math. Contr. Signals Syst.*, vol. 1, pp. 13-30, 1988.
- [3] ———, "On the supremal controllable sublanguage of a given language," *SIAM J. Contr. Optim.*, vol. 25, no. 3, pp. 637-659, May 1987.
- [4] Y. Brave and M. Heymann, "On stabilization of discrete-event processes," *Int. J. Contr.*, vol. 51, no. 5, pp. 1101-1117, 1990.
- [5] C. M. Özveren, A. S. Willsky, and P. J. Antaklis, "Stability and stabilizability of discrete-event dynamic systems," *J. ACM*, vol. 38, no. 3, pp. 730-752, 1991.
- [6] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages and Computations*. Reading, MA: Addison-Wesley, 1979.
- [7] S. Even, *Graph Algorithms*. Maryland: Computer Science Press, 1979.
- [8] Y. Brave and M. Heymann, "On optimal attraction in discrete-event processes," *Inform. Sci.*, vol. 67, pp. 245-267, 1993.
- [9] R. Kumar, V. Garg, and S. I. Marcus, "Stability of DES behavior," in *Proc. 1991 IFAC Intl. Symp. Distributed Intelligence Syst.*, Arlington, VA., 1991, pp. 13-18.
- [10] S. Eilenberg, *Automata, Languages, and Machines*, vol. A. New York: Academic, 1974.
- [11] S. LaFortune and E. Chen, "The infimal closed controllable superlanguage and its application in supervisory control," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 398-405, 1990.
- [12] F. Lin and W. M. Wonham, "On the computation of supremal controllable sublanguages," in *Proc. 23rd Annual Allerton Conf. Communication, Contr., Computing*, Urbana, IL, 1985, pp. 942-950.
- [13] M. Heymann, "Some algorithmic questions in discrete-event control," to appear.
- [14] H. Cho and S. I. Marcus, "On supremal languages of class of sublanguages that arise in synthesis problem with partial observations," *Math. Contr. Sig. Syst.*, vol. 2, pp. 47-69, 1989.
- [15] Y. Willner and M. Heymann, "On language convergence in discrete-event systems," in *Proc. 17th Conv. IEEE Israel*, Mar. 1991.
- [16] R. Kumar, V. Garg, and S. I. Marcus, "On language stability of DFDS," in *Proc. Int. Conf. Mathematical Theory Contr.*, I.I.T. Bombay, Bombay, India, 1990.
- [17] Y. Willner and M. Heymann, "Optimal language convergence in discrete-event control," to appear.
- [18] R. Kumar, V. Garg, and S. I. Marcus, "Language stability and stabilizability of discrete-event dynamical systems," *SIAM J. Contr. Optim.*, vol. 31, no. 5, pp. 1294-1320, 1993.
- [19] C. M. Özveren and A. S. Willsky, "Tracking and restrictability in discrete-event dynamical systems," *SIAM J. Contr. Optim.*, vol. 30, no. 6, pp. 1423-1446, 1992.



Yosef M. Willner received the B.Sc., the M.Sc., and the D.Sc. degrees in electrical engineering from Technion—Israel Institute of Technology, Haifa, Israel, in 1984, 1987, and 1992, respectively.

Since 1992, Dr. Willner has been a Research Associate in the Department of Education in Technology and Science, Technion. From 1993–1994, he was on leave at NASA—Ames Research Center, Moffet Field, CA, as an NRC Research Associate. His research interests include multivariable systems, adaptive control, discrete-event systems, hybrid systems, and graph algorithms.



Michael Heymann received the B.Sc. and the M.Sc. degrees from Technion—Israel Institute of Technology, Haifa, Israel, in 1960 and 1962, respectively, and the Ph.D. degree from the University of Oklahoma, Norman, in 1965, all in chemical engineering.

From 1965–1966, he was on the faculty of the University of Oklahoma. From 1966–1968, he was with Mobil Research and Development Corp., researching control and systems theory. From 1968–1970, he was with the Ben-Gurion University of the Negev, Beer-Sheva, where he established and headed the Department of Chemical Engineering department. Since 1970, he has been with the Technion, where he is currently a Professor in the Department of Computer Science, holding the Carl Fecheimer Chair. He has previously been with the Department of Electrical Engineering and Chairman of the Department of Applied Mathematics. He held visiting positions at the University of Toronto, the University of Florida, the University of Eindhoven, Concordia University, CSIR, Yale University, the University of Bremen, and the University of Newcastle. From 1983–1984, 1988–1989, and during several summers he was an NRC-Senior Research Associate with NASA—Ames Research Center. His current research interests include discrete-event systems, hybrid systems, and the theory of concurrent processes.

Dr. Heymann is on the editorial board of *SIAM Journal of Control and Optimization*.

Concurrent Vector Discrete-Event Systems

Yong Li, *Member, IEEE*, and W. M. Wonham, *Fellow, IEEE*

Abstract—The vector discrete-event system (VDES) is a compact serial discrete-event system model, in which the system state is represented by a vector with integer components and the transitions by integer vector addition [13], [14]. Continuing the study of VDES, we introduce in this paper concurrent VDES, extending the base VDES model to capture strict concurrency, or possible simultaneity of events. We characterize the effect of strict concurrency on control, and show how to synthesize nondeterministic controllers allowing maximal concurrency of a controlled VDES plant.

I. INTRODUCTION

CONCURRENCY is a critical issue in designing computer operating systems, data base management systems, and distributed computer and communication networks [4], [7], [1]. It is one of the key characteristics of real-time systems in general [18]. Researchers in the computer and communication areas have recognized that, compared to serial systems (in which events occur serially), concurrent systems present special challenges. The simultaneous event occurrences in a concurrent system make it harder to enforce the orderly flow of events according to a specification of desired behavior.

This paper studies concurrency from a control-theoretic perspective. We adopt the vector discrete-event system (VDES) introduced in [13] and [14] as the base model for controlled plants, incorporate simultaneity of events, and study its effect on controller synthesis. We call the presence of simultaneous events strict concurrency or simply concurrency.

As background, we first define VDES and review some concepts of VDES control.

VDES is a compact serial discrete-event system model. The state of a VDES is coordinatized by state variables x_1, x_2, \dots, x_n ranging over the integers \mathcal{Z} . The state space is the set of n -dimensional integer vectors

$$\mathcal{X} := \left\{ X \begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{matrix} \mid x_i \in \mathcal{Z} \right\} = \mathcal{Z}^n.$$

The event set Σ is partitioned into disjoint subsets Σ_c and Σ_u , the controllable-event subset and uncontrollable-event subset, respectively. The state transition is described by a partial function

$$\delta: \Sigma \times \mathcal{X} \rightarrow \mathcal{X}: (\alpha, X) \mapsto X + E_\alpha \quad (1)$$

Manuscript received July 22, 1993; revised May 2, 1994. Paper recommended by Associate Editor, S. Lafontaine. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

Y. Li is with Bell-Northern Research Ltd., Ottawa, Ont. K1Y 4H7, Canada. W. M. Wonham is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ont. M5S 1A4, Canada.

IEEE Log Number 9408276.

where $E_\alpha \in \mathcal{X}$ is the displacement vector for α . For $X \in \mathcal{X}$ we write $X \geq 0$ if $x_i \geq 0$ ($i = 1, 2, \dots, n$). $\delta(\alpha, X)$ is defined, denoted by $\delta(\alpha, X)!$, if

$$X + E_\alpha \geq 0.$$

A VDES is formally defined as a 4-tuple

$$G = (\Sigma, \mathcal{X}, \delta, X_0) \quad (2)$$

where X_0 is the initial state.

A state-feedback controller for G is a function

$$f: \mathcal{X} \rightarrow \Gamma \subseteq 2^\Sigma \quad (3)$$

where Γ is the set of control patterns defined as

$$\Gamma := \{\gamma \mid \Sigma_u \subseteq \gamma \subseteq \Sigma\}. \quad (4)$$

An event α is enabled at a state X if $\alpha \in f(X)$; otherwise, α is disabled.

We describe system behavior by predicates. A predicate on the state space \mathcal{X} is a function $P: \mathcal{X} \rightarrow \{0, 1\}$. The predicate operators “ \neg ” (negation), “ \wedge ” (conjunction), and “ \vee ” (disjunction) are defined as follows

$$(\neg P)(X) = 1 \Leftrightarrow P(X) = 0$$

$$(P_1 \wedge P_2)(X) = 1 \Leftrightarrow P_1(X) = 1 \text{ and } P_2(X) = 1$$

$$P_1 \vee P_2 = \neg((\neg P_1) \wedge (\neg P_2)).$$

A partial order is induced on the set of predicates on \mathcal{X} according to

$$P_1 \preceq P_2 \Leftrightarrow P_1 \wedge P_2 = P_1.$$

In this paper we identify a predicate P on \mathcal{X} with the subset of \mathcal{X} induced by P

$$\{X \in \mathcal{X} \mid P(X) = 1\}.$$

We will employ the expressions “ $X \in P$ ” and “ $P(X) = 1$ ” interchangeably.

For each $\alpha \in \Sigma$ the predicate transformation M_α , the weakest liberal precondition [16], is defined for a predicate P on \mathcal{X} by

$$M_\alpha(P)(X) := \begin{cases} 1 & \text{if } \delta(\alpha, X)! \text{ and } \delta(\alpha, X) \in P, \\ & \text{or if not } \delta(\alpha, X)! \\ 0 & \text{otherwise.} \end{cases}$$

For a predicate P on \mathcal{X} , we define a predicate $R(G, P)$ which characterizes the set of states reachable from X_0 by way of states satisfying P :

- 1) if $X_0 \in P$, then $X_0 \in R(G, P)$;
 - 2) if $X \in R(G, P)$, $\alpha \in \Sigma$, $\delta(\alpha, X)!$ and $\delta(\alpha, X) \in P$, then $\delta(\alpha, X) \in R(G, P)$;
 - 3) every state in $R(G, P)$ is obtained as in 1) and 2).
- We call a predicate P on \mathcal{X} controllable with respect to G if

$$P \preceq R(G, P) \quad \text{and} \quad (\forall \alpha \in \Sigma_u) P \preceq M_\alpha(P). \quad (5)$$

Note that the empty (or identically "false") predicate is trivially controllable.

For a VDES plant G under the control of a controller f , we define the closed-loop predicate $R(f/G)$ as follows [13]:

- 1) $X_0 \in R(f/G)$;
- 2) if $X \in R(f/G)$, $\alpha \in \Sigma$, $\delta(\alpha, X)!$ and $\alpha \in f(X)$, then $\delta(\alpha, X) \in R(f/G)$;
- 3) every state in $R(f/G)$ is obtained as in 1) and 2).

It was shown in [13] that a predicate P can be synthesized by a deterministic controller f for a serial VDES model G (i.e., $R(f/G) = P$) if and only if P is nonempty and controllable.

A controller f is called balanced [13] if for any $X, X' \in R(f/G)$ and any $\alpha \in \Sigma$ with $\delta(\alpha, X)!$ and $\delta(\alpha, X) = X'$ we have that $\alpha \in f(X)$. A balanced controller is the one which, among controllers synthesizing the same closed-loop predicate, enables the most events at any reachable state. It can be checked that any controller may be replaced by a balanced controller without changing the reachable set [9, A.2]. By concentrating on balanced controllers, we gain technical convenience in our proofs without loss of generality.

In the serial VDES model, the event-occurrence condition is defined for single events (in Σ) only; it is assumed that at any given instant of time, at most one event can occur. Also, at each state, the controller (3) issues a predetermined control pattern; such a controller is deterministic.

In this paper we first extend the serial VDES (2) to include concurrency (that is, to allow simultaneous occurrences of events); this extended VDES model is called concurrent VDES (CVDES) (Section II). Then we study the control of CVDES by both deterministic (Section III) and nondeterministic (Sections IV and V) controllers. A nondeterministic controller issues, at each state, a control pattern which is "randomly" selected from a set of control patterns. It is shown that nondeterministic controllers are more powerful than their deterministic counterparts in the presence of concurrency. We also discuss how to synthesize nondeterministic controllers which allow the controlled CVDES to operate with maximal concurrency according to a specified criterion.

In the control literature, simultaneous events in DES (but not VDES, which had not yet been introduced) were treated in a preliminary way in the thesis [8], via the ideas of well-posed supervisor, well-posed language, and concurrency product. It was shown that the supremal well-posed sublanguage of a given language need not exist, hence there need not exist an optimal supervisor in the presence of simultaneous events. These results were reported in [10] and [11], where nondeterminism was introduced to obtain an optimal solution. The completion of a set of state-feedback controls (SFBC), forming a poset but not a lattice, using nondeterminism (exponentiation) to construct a supremal element, was proposed

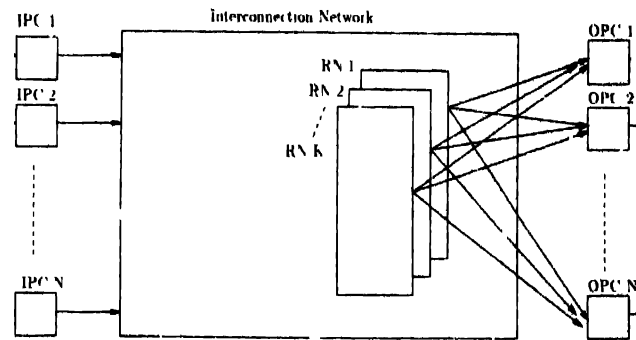


Fig. An ATM model.

independently in [2]. In [12] the optimal control problem for VDES with strict concurrency was formalized in the style of [10], [2], and [11], and the optimal nondeterministic SFBC obtained. The current paper completes [12] with full proofs of the authors' earlier results. We remark that the nonexistence, in the presence of strict concurrency, of a supremal deterministic SFBC, as distinct from merely multiple, maximal SFBC, was pointed out by Krogh for controlled Petri nets (CPN), when the latter were introduced in [5]. The existence of a unique maximal nondeterministic feedback policy, in the context of the forbidden state problem for CPN, was later noted in [3] and [6].

Throughout this paper, we use the example of an asynchronous transfer mode (ATM) switch. A general high-level model of an ATM switch [15] is shown in Fig. 1. It consists of N input-port controllers (IPC), an interconnection network and N output-port controllers (OPC). IPC and OPC are composed of buffers for incoming and outgoing ATM cells,¹ respectively, and the control logic for these buffers. The interconnection network is responsible for switching incoming cells to their destination output ports. The outgoing cells to an OPC arrive from K routing networks (RN). The buffer within an OPC has capacity C' ; while the buffer is full, any arriving cell will be discarded. An OPC can disable the transmission of cells from a RN by backward signaling.

In this paper we focus on a single OPC, modelling the buffer within the OPC and synthesizing control strategies for the buffer. For definiteness, assume that $K = 3$ and $C' = 5$.

We first model the buffer as the one-dimensional serial VDES G shown by its Petri-net graph in Fig. 2, with $E_{\alpha_1} = E_{\alpha_2} = E_{\alpha_3} = 1$, $E_{\alpha_4} = -1$, and $X_0 = 0$. In this model, α_i ($i = 1, 2, 3$) are controllable and α_4 is uncontrollable. The state transition graph of the buffer is displayed in Fig. 3. The control specification is that the buffer must not overflow, i.e., $(X \leq 5)$. Then the (optimal) controller enforcing this specification in G is easy to obtain

$$f^*(X) := \begin{cases} \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\} & \text{if } X < 5 \\ \{\alpha_4\} & \text{otherwise.} \end{cases} \quad (6)$$

In the serial model G , no two events can occur simultaneously, e.g., cells from two RN's never arrive at the OPC at the same time. This assumption is not realistic, considering the independence and high-rate of transmission of cells from

¹ An ATM cell is a segment of a 53-byte data stream.

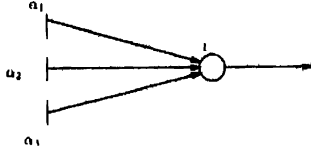


Fig. 2. The VDES model for an OPC buffer.

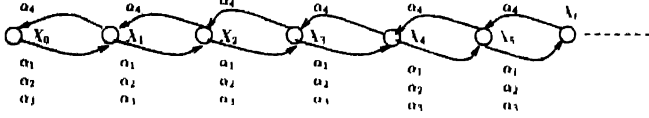


Fig. 3. The serial state transition graph of an OPC buffer

RN's in an ATM switch. In this paper we will show how to incorporate simultaneity into the VDES model of the OPC buffer and how to synthesize controllers in the presence of simultaneous events.

II. CONCURRENCY

We formalize concurrency by extending a VDES G to include simultaneous events. Define the extended event set by

$$\Sigma_{con} := 2^\Sigma.$$

An event $e = \langle \alpha_1, \alpha_2, \dots, \alpha_k \rangle \in \Sigma_{con}$ represents the simultaneous occurrence of $\alpha_1, \alpha_2, \dots, \alpha_k$. We define a simultaneous event as an unordered set: all its component events are thought of as occurring at the same instant of time, with no causal ordering among them. We reserve the symbol " $\langle \dots \rangle$ " for simultaneous events simply to differentiate them from ordinary (uninterpreted) subsets of Σ ; apart from this, " $\langle \dots \rangle$ " and " $\{ \dots \}$ " have the same mathematical meaning as subsets. In particular, $\emptyset \in \Sigma_{con}$ is the null simultaneous event, denoting that no event in Σ occurs. Our definition of simultaneous event rules out multiple simultaneous occurrences of the same event, but our results could easily be extended to accommodate that situation.

For $\alpha \in \Sigma$, $\langle \alpha \rangle \in \Sigma_{con}$ is a singleton event; with slight abuse of notation, we write $\langle \alpha \rangle \in \Sigma_{con}$ as α whenever no confusion results, as they really represent the same physical event.

Define a companion vector J_α for E_α

$$J_\alpha(i) := \begin{cases} -E_\alpha(i) & \text{if } E_\alpha(i) < 0 \\ 0 & \text{otherwise.} \end{cases} \quad (i = 1, 2, \dots, n) \quad (7)$$

We can regard $J_\alpha(i)$ as the minimal value of x_i required for the occurrence of α . Then define the occurrence condition of the simultaneous event e as

$$\delta_{con}(e, X)! \Leftrightarrow X \geq \sum_{\alpha \in e} J_\alpha. \quad (8)$$

This condition ensures that the resources represented by state variables are sufficient to support the simultaneous occurrence of all component events in e . We define the displacement vector for e as

$$E_e := \sum_{\alpha \in e} E_\alpha$$

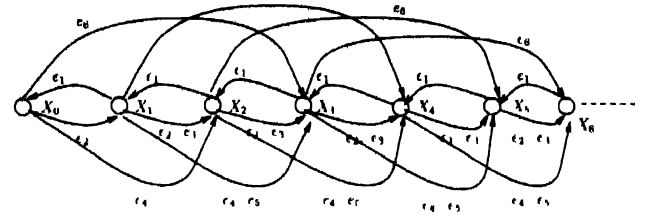


Fig. 4. The concurrent state transition graph of an OPC buffer.

and extend the state transition function as

$$\delta_{con}: \Sigma_{con} \times \mathcal{X} \rightarrow \mathcal{X}: (e, X) \mapsto X + E_e. \quad (9)$$

Note that δ_{con} is an extension of δ in (1) in the sense that

$$\delta_{con}(\alpha, X) \mapsto \delta(\alpha, X).$$

On this understanding, from now on we drop the subscript "con" from δ_{con} and write simply δ . The extended model is called concurrent VDES (CVDES) and is denoted by

$$G_{con} := (\Sigma_{con}, \mathcal{X}, \delta, X_0). \quad (10)$$

As an illustration, consider the one-dimensional VDES model for the OPC buffer in Fig. 2. The extended event set $\Sigma_{con} = 2^\Sigma$ with $\Sigma := \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$. For instance, $e := \langle \alpha_1, \alpha_2, \alpha_4 \rangle \in \Sigma_{con}$ denotes the event that two cells from RN 1 and RN 2 arrive at the OPC at the same time as a cell is transmitted from the OPC. The simultaneous event e is defined at every state except $X = 0$ since

$$\sum_{\alpha \in e} J_\alpha = 1.$$

The state transition graph of G_{con} is shown in Fig. 4, where e_1 denotes α_4 ; e_2 denotes the events

$$\alpha_i \quad (i = 1, 2, 3)$$

e_3 denotes

$$\langle \alpha_i, \alpha_j, \alpha_4 \rangle \quad (i, j = 1, 2, 3, i \neq j)$$

e_4 denotes the events

$$\langle \alpha_i, \alpha_j \rangle \quad (i, j = 1, 2, 3; i \neq j)$$

e_5 denotes

$$\langle \alpha_1, \alpha_2, \alpha_3, \alpha_4 \rangle$$

and e_6 denotes $\langle \alpha_1, \alpha_2, \alpha_3 \rangle$.

The following simple properties of CVDES will be used later.

Claim 1: Let $e \in \Sigma_{con}$, and $e_1 \subseteq e$, $e_2 := e - e_1$. If $\delta(e, X)!$ then $X_1 := \delta(e_1, X)!$ and $\delta(e_2, X_1)!$, and then $\delta(e, X) = \delta(e_2, X_1)$.

Proof: Let $e_1 \subseteq e$ and $\delta(e, X)!$. Then

$$X \geq \sum_{\alpha \in e} J_\alpha \geq \sum_{\alpha \in e_1} J_\alpha$$

with J_α being the companion vector E_α as defined in (7). So, $\delta(e_1, X)!$, according to (8). From (7)

$$(\forall \alpha) E_\alpha \geq -J_\alpha.$$

Thus

$$X_1 := \delta(e_1, X) = X + \sum_{\alpha \in e_1} E_\alpha \geq X + \sum_{\alpha \in e_1} (-J_\alpha) \geq \sum_{\alpha \in e_2} J_\alpha$$

since

$$X \geq \sum_{\alpha \in e} J_\alpha = \sum_{\alpha \in e_1} J_\alpha + \sum_{\alpha \in e_2} J_\alpha.$$

So, $\delta(e_2, X_1)!$, as required. By the definition, $\delta(e, X) = \delta(e_2, X_1)$. \square

We call $w \in \Sigma^*$ an instance event string of the simultaneous event $e \in \Sigma_{con}$ if each element of e occurs exactly once in w . For example, $w = \alpha_1 \alpha_2 \alpha_3$ is an instance event string of $e = \langle \alpha_1, \alpha_2, \alpha_3 \rangle$.

Claim 2: Let w be an instance event string of $e \in \Sigma_{con}$. Then for all $X \in \mathcal{X}$, $\delta(e, X)!$ only if $\delta(w, X)!$, and then $\delta(w, X) = \delta(e, X)$.

Proof: By the definition of δ . \square

Claim 3: Let $S \subseteq \Sigma$ and P be a predicate which is controllable wrt G . Assume $X \in P$. Then

$$\begin{aligned} (\forall e \in 2^S)(\delta(e, X) \in P \text{ or not } \delta(e, X)!) \\ \Leftrightarrow (\forall e \in 2^{\Sigma_{con}})(\delta(e, X) \in P \text{ or not } \delta(e, X)!) \end{aligned}$$

Proof: (\Rightarrow) Automatic. \square

(\Leftarrow) Let $e \in 2^S$ and $\delta(e, X)!$. To show that $\delta(e, X) \in P$, let $e_1 := e \cap \Sigma_c$, $e_2 := e - e_1$. By Claim 1

$$X_1 := \delta(e_1, X)!, \quad \delta(e_2, X_1)!$$

and

$$\delta(e, X) = \delta(e_2, X_1).$$

By hypothesis, $X_1 \in P$. Let $w \in \Sigma_u^*$ be an instance event string of e_2 . Since P is controllable, by Claim 2

$$\delta(e_2, X_1) = \delta(w, X_1) \in P$$

namely $\delta(e, X) \in P$, as claimed. \square

For CVDES, a controller controls singleton events in Σ_c directly; a simultaneous event can be controlled if and only if at least one of its component events is represented in Σ_c . Thus the uncontrollable events in G_{con} are $\Sigma_{con,u} := 2^{\Sigma_u}$ (so, $\emptyset \in \Sigma_{con,u}$), and the controllable events are $\Sigma_{con,c} := \Sigma_{con} - \Sigma_{con,u}$.

Moreover, a simultaneous event is disabled by a controller if and only if one or more of its component events is disabled; or equivalently, a simultaneous event is enabled if and only if all of its component events are enabled. We formalize this requirement by defining Γ_{con} , the set of admissible control patterns of G_{con} . A control pattern of G_{con} will be an element

$\eta \in 2^{\Sigma_{con}}$. Thus each η is a set of simultaneous events, interpreted as enabled in Σ_{con} . Formally

$$\Gamma_{con} := \{\eta \mid \eta = 2^\gamma, \Sigma_u \subseteq \gamma \subseteq \Sigma\} \quad (11)$$

or equivalently

$$\Gamma_{con} := \{\eta \mid \eta = 2^\gamma, \gamma \in \Gamma\}.$$

Therefore, each control pattern $\eta \in \Gamma_{con}$ contains among its singleton events all uncontrollable events in Σ_u ; and η is generated by its singleton events. Denote the set of singleton events in a control pattern $\eta \in \Gamma_{con}$ by $\bar{\eta}$. Then $\eta = 2^{\bar{\eta}}$; we call this property of control patterns control dependency. Note that the null simultaneous event $\emptyset \in \Sigma_{con}$ is included in each control pattern η . This can be interpreted to mean that “no event occurs” is always allowed.

For the OPC buffer example, $\Sigma_{con,u} = \{\emptyset, \alpha_4\}$ and a control pattern in Γ_{con} takes one of the forms

$$\{\emptyset, \alpha_4\}$$

$$\{\emptyset, \alpha_4, \alpha_i, e_{i,4}\} \quad i = 1, 2, 3$$

$$\{\emptyset, \alpha_4, \alpha_i, \alpha_j, e_{i,4}, e_{j,4}, e_{i,j}, e_{i,j,4}\} \quad i, j = 1, 2, 3: i \neq j$$

$$\{\emptyset, \alpha_4, \alpha_1, \alpha_2, \alpha_3, e_{1,4}, e_{2,4}, e_{3,4}, e_{1,2}, e_{1,3}, e_{2,3}, e_{1,2,3}, e_{1,2,4}, e_{2,3,4}, e_{1,2,3,4}\}$$

where $e_{i,j}, e_{i,j,k}, e_{1,2,3,4}$ denote $\langle \alpha_i, \alpha_j \rangle, \langle \alpha_i, \alpha_j, \alpha_k \rangle, \langle \alpha_1, \alpha_2, \alpha_3, \alpha_4 \rangle$, respectively.

Note that a simultaneous event is included in a control pattern above if and only if its component events are included: In the case of the simultaneous event, say $e_{1,2}$, this means that two cells from RN 1 and RN 2 are allowed to arrive at the OPC at the same time instant if (and only if) these two RN's are both enabled for transmission by the OPC at that time.

Define a controller for G_{con} as $f': \mathcal{X} \rightarrow \Gamma_{con}$. Let $R(f'/G_{con})$ be the closed-loop predicate (i.e., the set of reachable states in the closed-loop system) synthesized by f' in G_{con} .²

A question of interest is the following.

Q_1 . For a given predicate P on \mathcal{X} , how to synthesize a controller f' such that $R(f'/G_{con}) = P$?

When the serial model G is extended to G_{con} to include concurrency, the control action of a controller $f: \mathcal{X} \rightarrow \Gamma$ for G is extended accordingly and is captured by a controller f_{con} for G_{con}

$$(\forall X \in \mathcal{X}) f_{con}(X) := 2^{f(X)}.$$

At a state X , f_{con} enables all singleton events enabled by f at X , as well as all simultaneous events generated by these singleton events. We call this controller the concurrent extension of f . Conversely, for any $f': \mathcal{X} \rightarrow \Gamma_{con}$, if we define a controller $f: \mathcal{X} \rightarrow \Gamma$ by

$$(\forall X \in \mathcal{X}) f(X) := \bar{f}'(X)$$

²The formal definition of $R(f'/G_{con})$ can be given in the same way as that of $R(f/G)$ in Section 1.

where $\bar{f}'(X)$ denotes the singleton-event set of $f'(X)$, then $f_{con} = f'$ since

$$(\forall X \in \mathcal{X}) f'(X) = 2^{\bar{f}'(X)}.$$

This controller f is called the serial reduction of f' . We say that f' is balanced if its serial reduction is.

It should be pointed out that, in implementation, the physical control action of a controller f_{con} is identical to that of its serial reduction f ; both directly enable the same set of singleton events. But, owing to control dependence, f_{con} also implicitly enables all simultaneous events generated by enabled singleton events. As far as the closed-loop predicate is concerned, f_{con} is more permissive for G_{con} than f is for G , simply because the former allows more events to occur at any state than the latter. We say that concurrency has no effect on the control action of f if $R(f/G) = R(f_{con}/G_{con})$; in this case, the synthesis of f_{con} for G_{con} is reduced to that of f for G .

Another question of interest is then the following.

Q_2 . Under what condition does concurrency have no effect on the control action of f ?

We address Q_1 and Q_2 in the next section.

III. CONCURRENT WELL-POSEDNESS

We have seen in Section I that the balanced controller (6) synthesizes the controllable specification $P = (X \leq 5)$ in the serial model G for the OPC buffer. We now show that there is no balanced controller which synthesizes P in the concurrent model G_{con} (Fig. 4). Suppose otherwise, and let f' be such a controller. Then $X_1, X_5 \in R(f'/G_{con})$. Since f' is balanced, we have that $\alpha_i \in f'(X_i)$ ($i = 1, 2, 3$). By control dependence, $e_{1,2,3} \in f'(X_1)$. But $\delta(e_{1,2,3}, X) = 4 + 3 = 7 \notin P$, a contradiction.

This shows that controllability alone is no longer a sufficient condition for a predicate to be synthesized as a closed-loop predicate for a CVDES plant; and that $R(f/G) = R(f_{con}/G_{con})$ is not always true.

A predicate P is concurrently well-posed (CWP) wrt G_{con} if for any $X \in P$ and any $\alpha, \beta \in \Sigma$ with $\delta(\alpha, X) \in P$, $\delta(\beta, X) \in P$ and $\delta(\langle \alpha, \beta \rangle, X)!$, we have $\delta(\langle \alpha, \beta \rangle, X) \in P$.

It can be easily checked that the specification P for the ATM example is not CWP.

The following theorem answers questions Q_1 and Q_2 posed in the previous section, and asserts that CWP together with controllability is a necessary and sufficient condition to guarantee that there exists a (balanced) controller synthesizing P in G_{con} ; and that when the closed-loop predicate synthesized by a controller for G is CWP, concurrency has no effect on its control action.

Theorem 4: 1) There exists a balanced controller $f': \mathcal{X} \rightarrow \Gamma_{con}$ such that

$$R(f'/G_{con}) = P$$

if and only if P is nonempty, controllable, and CWP.

2) Let $f: \mathcal{X} \rightarrow \Gamma$ be a balanced controller for G and f_{con} be its concurrent extension. Then

$$R(f/G) = R(f_{con}/G_{con})$$

if and only if $R(f/G)$ is CWP.

Proof: 1) (IF) Let the predicate P be nonempty, controllable, and CWP. Define a controller f by

$$(\forall \alpha \in \Sigma_c) f_\alpha(X) := \begin{cases} 1 & \text{if } \delta(\alpha, X)! \text{ and } \delta(\alpha, X) \in P \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

Then as shown in the proof of [13] Theorem 6, $R(f/G) = P$. It is easy to check that f thus defined is balanced. It remains to show that $R(f/G) = R(f_{con}/G_{con})$. We only need to show that $R(f_{con}/G_{con}) \preceq R(f/G)$, as the reverse inequality is automatic. First note that $X_0 \in R(f/G)$, $X_0 \in R(f_{con}/G_{con})$. For induction on the length of event strings in Σ_{con}^* , suppose that $X \in R(f/G)$, $X \in R(f_{con}/G_{con})$. Let $e \in \Sigma_{con}$, $\delta(e, X)!$ and $e \in f_{con}(X)$ (i.e., $(\forall \alpha \in e) \alpha \in f(X)$). We wish to show that

$$\delta(e, X) \in R(f/G). \quad (13)$$

The proof is again by induction, this time on the cardinality of e .

When $|e| = 1$, $e = \alpha$ for some $\alpha \in \Sigma$. In this case, (13) is true since $\alpha \in f_{con}(X)$ implies $\alpha \in f(X)$.

When $|e| = 2$, $e = \langle \alpha_1, \alpha_2 \rangle$ for some $\alpha_1, \alpha_2 \in \Sigma$. Since $\langle \alpha_1, \alpha_2 \rangle \in f_{con}(X)$, $\alpha_1, \alpha_2 \in f(X)$ and

$$\delta(\alpha_1, X), \delta(\alpha_2, X) \in R(f/G) = P.$$

But P is CWP by assumption, so

$$\delta(\langle \alpha_1, \alpha_2 \rangle, X) \in P = R(f/G).$$

Now, suppose that (13) is true for $|e| \leq k+1$. We show that it is also true for $|e| = k+2$. Let $e = e' \cup \langle \beta_1, \beta_2 \rangle$ with $|e'| = k$. Since $\delta(e, X)!$ we know from Claim 1 that $e', e - \langle \beta_1 \rangle$ and $e - \langle \beta_2 \rangle$ are all defined at X . It is easy to see that, when e is enabled by f_{con} at X , so are $e', e' \cup \langle \beta_1 \rangle$ and $e' \cup \langle \beta_2 \rangle$. We then have by the induction hypothesis

$$\delta(e', X) \in R(f/G)$$

$$\delta(e' \cup \langle \beta_1 \rangle, X) \in R(f/G)$$

$$\delta(e' \cup \langle \beta_2 \rangle, X) \in R(f/G)$$

that is

$$\delta(\langle \beta_1 \rangle, \delta(e', X)) \in R(f/G)$$

$$\delta(\langle \beta_2 \rangle, \delta(e', X)) \in R(f/G).$$

Again, since $P (= R(f/G))$ is CWP it follows that

$$\delta(e' \cup \langle \beta_1, \beta_2 \rangle, X) = \delta(\langle \beta_1, \beta_2 \rangle, \delta(e', X)) \in R(f/G)$$

if we notice that $\delta(\langle \beta_1, \beta_2 \rangle, \delta(e', X))!$ (Claim 1).

(ONLY IF) Let

$$R(f_{con}/G_{con}) = P$$

for a balanced controller f_{con} . By an argument similar to the proof of [13] Theorem 6, we can check that P is controllable. It remains to show that P is CWP. Let $X \in P$ and $\delta(\langle \alpha, \beta \rangle, X)!$ with

$$\delta(\alpha, X) \in P, \quad \delta(\beta, X) \in P.$$

Then

$$X, \delta(\alpha, X), \delta(\beta, X) \in R(f_{con}/G_{con}).$$

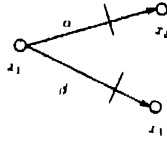


Fig. 5. A VDES example.

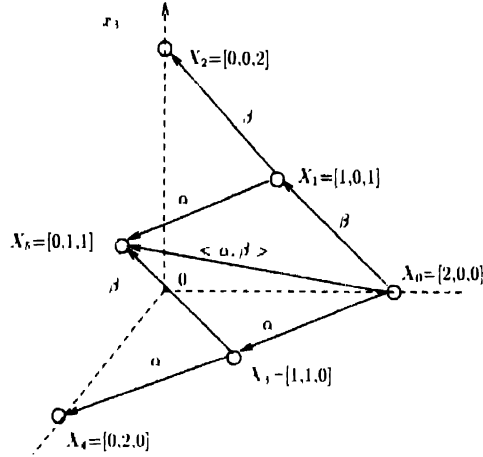


Fig. 6. The state transition graph of the VDES example.

But f_{con} is balanced, so

$$\alpha, \beta \in f_{con}(X)$$

and therefore, by control dependence

$$\langle \alpha, \beta \rangle \in f_{con}(X).$$

We then have

$$\delta(\langle \alpha, \beta \rangle, X) \in R(f_{con}/G_{con}) = P$$

and that P is CWP.

2) Easily follows from 1). \square

The above result indicates that when P is both controllable and CWP the construction of a balanced controller f_{con} synthesizing P in G_{con} is reduced to that of a controller f synthesizing P in G ; such a controller f can be constructed as in (12).

If there exists a supremal element in the set of all controllable and CWP subpredicates of a given predicate specifying legal states, an optimal solution can be found to the deterministic-controller synthesis problem for G_{con} . Unfortunately, such a supremal element may not exist. To see this, consider the Petri-net graph G in Fig. 5, with both α and β controllable. The state transition graph for G_{con} is shown in Fig. 6. Let

$$P_1 := \{X_0, X_1\}$$

$$P_2 := \{X_0, X_3\}$$

$$P := P_1 \vee P_2.$$

It can be easily checked in Fig. 6 that both P_1 and P_2 are controllable and CWP. But P is not CWP (although it is controllable), since $\delta(\alpha, X_0) = X_3 \in P$, $\delta(\beta, X_0) = X_1 \in P$, but $\delta(\langle \alpha, \beta \rangle, X_0) = X_5 \notin P$.

This shows that optimal deterministic control is not always possible,³ revealing the irregular nature of concurrency in our model. However, as we show in the next section, this irregularity is present only for deterministic control; optimality is always achievable with nondeterministic control.

IV. NONDETERMINISTIC CONTROL WITH CONCURRENCY

In the previous section we showed that there is no balanced deterministic controller synthesizing the specification P in the concurrent model G_{con} of the OPC buffer. Even though we can find unbalanced deterministic controllers synthesizing P in G_{con} , these controllers could be too restrictive to be of practical use. For instance, the OPC which permanently enables α_1 and disables α_2 and α_3 (i.e., only RN 1 is allowed to transmit cells to the OPC) will synthesize P in G_{con} , but this controller is obviously undesirable. In fact, owing to control dependence, any unbalanced deterministic controller synthesizing P will, at the state X_4 , enable one predetermined RN and disable the other two; and at X_3 , enable one or two predetermined RN's and disable the rest. This type of predetermined preference shown by the deterministic controller is undesirable, if not unacceptable, for the OPC design. (Suppose RN 1 is enabled and RN 2, RN 3 are disabled at X_1 . Then it is possible that the OPC oscillates between the states X_4 and X_5 , denying transmission by RN 2 and RN 3 indefinitely.) However, this restriction can be removed by use of nondeterministic control. For instance, instead of enabling a predetermined controllable event we can randomly⁴ enable one of the three controllable events whenever the state X_4 is reached. Similarly, we can randomly enable two controllable events at the state X_3 . The actual random selection mechanism might depend on, say, relative sizes of queues for the waiting cells in the three routing networks; but we omit details of implementation here.

The above discussion indicates that nondeterministic control can be more powerful than its deterministic counterpart in the presence of concurrency. In the following, we formalize nondeterministic control and exploit its power.

A nondeterministic controller (NDC) for G is a function $f_n: X \rightarrow 2^I$. A nondeterministic controller (NDC) for G_{con} is a function $f'_n: X \rightarrow 2^{\Gamma_{con}}$. Intuitively, the control pattern issued by an NDC f_n (or, f'_n) at X is a randomly chosen element among the elements of $f_n(X) \subseteq I$ (or, $f'_n(X) \subseteq \Gamma_{con}$).

Subject to concurrency in the plant (modeled by G_{con}), a (physical) NDC modeled by f_n will behave like a (physical) NDC modeled by $f_{n,con}$ defined as

$$(\forall X \in X) f_{n,con}(X) := \{2^\gamma \mid \gamma \in f_n(X)\}$$

because of control dependence. We call $f_{n,con}$ the concurrent extension of f_n . Conversely, for any NDC f'_n for G_{con} , if we

³One special case where the optimal controller exists for a specification P in G_{con} is when $\text{sup}C(P)$ turns out to be CWP. In this case, $\text{sup}C(P)$ is also the supremal element of the set of controllable and CWP subpredicates of P . Then we can construct an optimal controller f'_{con} such that $R(f'_{con}/G_{con}) = \text{sup}C(P)$.

⁴Here "randomly" just means "according to some hidden or unmodeled mechanism of selection."

define an NDC f_n for G as

$$(\forall X \in \mathcal{X}) f_n(X) := \{\bar{\eta} \mid \eta \in f'_n(X)\}$$

where $\bar{\eta}$ denotes the set of singleton events of η , then $f'_n = f_{n,con}$. We call f_n the serial reduction of f'_n . It should be emphasized that, in implementation, the physical control actions of f_n and $f_{n,con}$ are the same; they directly enable at any state a pattern γ of singleton events selected from the same subset of Γ . But, owing to control dependence, $f_{n,con}$ implicitly also enables at that state all simultaneous events generated by γ .

For a VDES plant G under the control of the NDC f_n , we define the closed-loop predicate $R(f_n/G)$ as follows:

- 1) $X_0 \in R(f_n/G)$;
- 2) if $X \in R(f_n/G)$, $\alpha \in \Sigma$, $\delta(\alpha, X)!$ and $(\exists \gamma \in f_n(X)) \alpha \in \gamma$, then $\delta(\alpha, X) \in R(f_n/G)$;
- 3) every state in $R(f_n/G)$ is obtained as in 1) and 2).

Similarly, we can define the closed-loop predicate $R(f'_n/G_{con})$.

To extend the definition of balanced deterministic controllers, we call an NDC f_n for G balanced if the following two technical conditions hold.

- 1) For any $X, X' \in R(f_n/G)$ and $\alpha \in \Sigma$ with $\delta(\alpha, X)!$ and $X' = \delta(\alpha, X)$, there exists a control pattern $\gamma \in f_n(X)$ such that $\alpha \in \gamma$.

- 2) There is no redundant control pattern; that is, for any state X and $\gamma_1, \gamma_2 \in f_n(X)$, if $\gamma_1 \subseteq \gamma_2$ then $\gamma_1 = \gamma_2$.

We call an NDC $f_{n,con}$ balanced if f_n is, and confine attention to balanced NDC.

Theorem 5: Let P be a predicate on \mathcal{X} with $X_0 \in P$. There exists a balanced NDC $f_n: \mathcal{X} \rightarrow 2^\Gamma$ such that

$$R(f_{n,con}/G_{con}) = P$$

if and only if P is controllable.

Proof: (IF) Assume that P is controllable with respect to G . Let $f: \mathcal{X} \rightarrow \Gamma$ be defined as in (12). Then $R(f/G) = P$. We define an NDC $f'_n: \mathcal{X} \rightarrow 2^\Gamma$ as

$$f'_n(X) = \{\Sigma_u \cup \{\alpha\} \mid \alpha \in f(X) \cap \Sigma_c\}. \quad (14)$$

Each control pattern γ in $f'_n(X)$ has the form $\Sigma_u \cup \{\alpha\}$, α being a controllable event which is enabled by f at X . We need to show that

$$R(f'_{n,con}/G_{con}) = P.$$

First, we show that

$$P \preceq R(f'_{n,con}/G_{con}).$$

By the assumption, $X_0 \in P$, and by the definition, $X_0 \in R(f'_{n,con}/G_{con})$. For a proof by induction, suppose that $X \in P$ with

$$X = \delta(\alpha, X')$$

for some $X' \in R(f/G) = P$ and $X' \in R(f'_{n,con}/G_{con})$. By the definition, $\alpha \in f(X')$, $\alpha \in \gamma$ for some $\gamma \in f'_n(X')$, and $\langle \alpha \rangle \in \eta$ for some $\eta \in f'_{n,con}(X')$. By the induction hypothesis

$$X \in R(f'_{n,con}/G_{con}).$$

This proves $P \preceq R(f'_{n,con}/G_{con})$. Next we prove the reverse inequality. Again, by the assumption, $X_0 \in P$, and by the definition, $X_0 \in R(f'_{n,con}/G_{con})$. For a proof by induction, suppose that

$$X \in R(f'_{n,con}/G_{con})$$

with

$$X = \delta(e, X'), \quad X' \in R(f'_{n,con}/G_{con}), \quad X' \in P$$

and

$$e \in \eta \text{ for some } \eta \in f'_{n,con}(X').$$

We show that $X \in P$. By the definition of $f'_{n,con}$

$$e \subseteq \Sigma_u \cup \{\alpha\}$$

for some α satisfying

$$\alpha \in \Sigma_c \cap f(X').$$

Case 1: $\alpha \notin e$. In this case, $e \subseteq \Sigma_u$. Since P is controllable with respect to G

$$\delta(w, X') \in P$$

with w any instance event string of e . So from Claim 2

$$X = \delta(e, X') = \delta(w, X') \in P$$

Case 2: $\alpha \in e$. Since $\alpha \in \Sigma_c \cap f(X')$, we have

$$\delta(\alpha, X') \in R(f/G) = P.$$

Let w be an instance event string of $e - \{\alpha\}$. Since $e - \{\alpha\} \subseteq \Sigma_u$, and P is controllable with respect to G , there follows

$$\delta(w, \delta(\alpha, X')) \in R(f/G) = P.$$

We then have from Claim 2 that

$$\delta(e, X') = \delta(w, \delta(\alpha, X')) \in P.$$

This completes the proof that

$$R(f'_{n,con}/G_{con}) \preceq P.$$

It is straightforward to check that f'_n is balanced.

(ONLY IF) Let $f_{n,con}$ be a balanced NDC. The definition of admissible control patterns implies that uncontrollable events are always enabled by $f_{n,con}$. It then follows easily that $R(f_{n,con}/G_{con})$ is controllable. \square .

The above result has two implications. One is that (balanced) NDC's synthesize the same set of closed-loop predicates in G_{con} as deterministic controllers in the serial model G : nondeterminism of controllers "offsets" the effect of concurrency on control. In other words, in contrast to deterministic control, the CWP condition is not required for a predicate to be synthesized by a nondeterministic controller; the NDC f'_n constructed in (14) always synthesizes a controllable predicate P . Since for any predicate P there exists a supremal controllable subpredicate of P [13], the second implication is that we can construct a balanced NDC f_n whose concurrent extension $f_{n,con}$ synthesizes the largest possible closed-loop subpredicate of P in G_{con} . In fact, there may be many

such controllers f_n . For the ATM example, the concurrent extensions of both f_n^1 and f_n^2 defined below synthesize the specification P in G_{con} .

$$f_n^1(X) := \{\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}\} \quad (X = 0, 1, 2)$$

$$f_n^1(X) := \{\{\alpha_1, \alpha_4\}, \{\alpha_2, \alpha_4\}, \{\alpha_3, \alpha_4\}\} \quad (X = 3, 4)$$

$$f_n^1(X) := \{\{\alpha_4\}\} \quad (X \geq 5)$$

$$f_n^2(X) := f_n^1(X) \quad (X \neq 3)$$

$$f_n^2(X) := \{\{\alpha_1, \alpha_2, \alpha_4\}, \{\alpha_1, \alpha_3, \alpha_4\}, \{\alpha_2, \alpha_3, \alpha_4\}\} \quad (X = 3).$$

The NDC f_n^2 differs from f_n^1 at the state $X = 3$ only, where it enables two controllable events instead of only one. Therefore, the concurrent extension of f_n^2 allows more concurrency than f_n^1 in the concurrent model G_{con} .

In the next section we investigate how to pick, among all NDC's synthesizing a given predicate, one which optimizes another performance measure—namely, concurrency.

V. NONDETERMINISTIC CONTROLLERS ALLOWING MAXIMUM CONCURRENCY

For the controlled CVDES G_{con} , concurrency is measured by the simultaneity allowed by a controller. A deterministic controller f_{con} allows greater concurrency of the plant than another deterministic controller g_{con} if

$$(\forall X \in \mathcal{X}) f_{con}(X) \supseteq g_{con}(X)$$

i.e., at any state there are more events enabled by f_{con} than by g_{con} .

We now introduce a measure of concurrency permitted to the controlled plant by NDC's. For a set S and two elements $A, B \in 2^{(2^S)}$, we write $A \leq B$ (wrt S) if

$$(\forall a \in A)(\exists b \in B)a \subseteq b.$$

For a predicate P define a relation " \leq " on the set of all NDC's of G by

$$f_n \leq g_n (rel P) \text{ if } (\forall X \in P) f_n(X) \leq g_n(X) \text{ (wrt } \Sigma).$$

Similarly, define for P a relation " \leq " on the set of NDC's of G_{con} by

$$f_{n,con} \leq g_{n,con} (rel P) \text{ if } (\forall X \in P) f_{n,con}(X) \leq g_{n,con}(X) \text{ (wrt } \Sigma_{con}).$$

The meaning of " \leq " is easy to see: $f_{n,con} \leq g_{n,con} (rel P)$ implies that at any state X of P , for any control pattern which is chosen by $f_{n,con}$ there is a control pattern chosen by $g_{n,con}$ which specifies more enabled simultaneous events. Therefore, the relation " \leq " on the set of NDC's for G_{con} imposes an ordering on the degree of permissiveness and concurrency they allow of the plant. It is easy to check that

$$f_n \leq g_n (rel P) \Leftrightarrow f_{n,con} \leq g_{n,con} (rel P).$$

The following simple result says that the ordering " \leq " on the set of NDC's induces a consistent ordering among the closed-loop predicates.

Claim 6: If $f_{n,con} \leq g_{n,con} (rel R(f_{n,con}/G_{con}))$ then

$$R(f_{n,con}/G_{con}) \preceq R(g_{n,con}/G_{con}).$$

□

For a controllable predicate P , let f_n^o be defined as in (14). Define

$$F^+(P) := \{f_n \mid f_n^o \leq f_n (rel P)\}.$$

$F^+(P)$ comprises all controllers which are more permissive in their control action than f_n^o . It follows from Claim 6 and the fact $R(f_{n,con}^o/G_{con}) = P$ that

$$(\forall f_n \in F^+(P)) P \preceq R(f_{n,con}/G_{con}). \quad (15)$$

For a controllable predicate P , we also define

$$F^-(P) := \{f_n \mid (\forall X \in P)(\forall \gamma \in f_n(X))(\forall e \in 2^\gamma) \cdot \delta(e, X) \in P \text{ or not } \delta(e, X)!\}.$$

$F^-(P)$ comprises all controllers whose concurrent extensions enable only legal simultaneous events. It follows that

$$(\forall f_n \in F^-(P)) R(f_{n,con}/G_{con}) \preceq P. \quad (16)$$

The next result says that the concurrent extensions of balanced controllers in $F^-(P) \cap F^+(P)$ are exactly the controllers which synthesize P in G_{con} ; $F^-(P)$ determines a lower range, and $F^+(P)$ an upper range, of such controllers.

Theorem 7: Assume that P is controllable and $X_0 \in P$. Then for any balanced controller $f_{n,con}$

$$R(f_{n,con}/G_{con}) = P \Leftrightarrow f_n \in F^-(P) \cap F^+(P).$$

Proof: (\Leftarrow) Directly follows from (15) and (16).

(\Rightarrow) Assume for a balanced controller $f_{n,con} : \mathcal{X} \rightarrow 2^{\Gamma_{con}}$

$$R(f_{n,con}/G_{con}) = P.$$

To show that $f_n \in F^+(P)$, suppose for a contradiction that

$$\text{not } (f_n^o \leq f_n) (rel P).$$

Then

$$(\exists X \in P) \text{ not } (f_n^o(X) \leq f_n(X))$$

that is

$$(\exists X \in P)(\exists \gamma \in f_n^o(X))(\forall \gamma_1 \in f_n(X)) \text{ not } (\gamma \subseteq \gamma_1).$$

Since all candidate control patterns in $f_n^o(X)$ have the form

$$\Sigma_u \cup \{\alpha\}$$

there follows

$$(\exists X \in P)(\exists \alpha \in \Sigma_c)(\Sigma_u \cup \{\alpha\} \in f_n^o(X) \& (\forall \gamma \in f_n(X)) \alpha \notin \gamma).$$

So

$$(\exists X \in P)(\exists \alpha \in \Sigma_c) \delta(\alpha, X)! \& \delta(\alpha, X) \in P \& (\forall \gamma \in f_n(X)) \alpha \notin \gamma.$$

But this contradicts the assumption that $f_{n,con}$ (or f_n) is balanced.

The conclusion that $f_n \in F^-(P)$ follows directly from the assumption that $R(f_{n,con}/G_{con}) = P$ and the definition of $F^-(P)$. \square

For a controllable predicate P , denote the balanced controllers in $F^+(P) \cap F^-(P)$ by $F(P)$.

Lemma 8: $F(P)$ is a poset with respect to the relation " \leq ".

Proof: It is easy to see that the relation " \leq " is both reflexive and transitive.

To show that it is also antisymmetric, let $f_n, g_n \in F(P)$ with

$$f_n \leq g_n, g_n \leq f_n \quad (rel P).$$

We show that $f_n = g_n (rel P)$. For any $X \in P$, let $\gamma \in f_n(X)$. Then $\gamma \subseteq \gamma_1$ for some $\gamma_1 \in g_n(X)$ since $f_n \leq g_n$. Also, since $g_n \leq f_n$, we have $\gamma_1 \subseteq \gamma_2$ for some $\gamma_2 \in f_n(X)$. So $\gamma \subseteq \gamma_1 \subseteq \gamma_2$. Since f_n is balanced, there is no redundant control pattern in $f_n(X)$. Thus $\gamma = \gamma_2$, implying that $\gamma = \gamma_1$. This proves that

$$(\forall \gamma \in f_n(X))(\exists \gamma_1 \in g_n(X))\gamma = \gamma_1$$

or, equivalently

$$f_n(X) \subseteq g_n(X).$$

Repeating the above argument with f and g interchanged, we get

$$f_n(X) \supseteq g_n(X).$$

So

$$(\forall X \in P)f_n(X) = g_n(X)$$

i.e., $f_n = g_n (rel P)$. \square

By the definitions of $F(P)$ and f_n^o , we know that f_n^o is the infimal element of the poset $F(P)$. We now construct the supremal element of $F(P)$.

For a controllable predicate P on \mathcal{X} and $X \in P$ define

$$M(X) := \{c \mid c \subseteq \Sigma_c \text{ and } (\forall c' \in 2^c)\delta(c', X) \in P \text{ or not } \delta(c', X)!\}$$

and

$$N(X) := \{\Sigma_u \cup c \mid c \in M(X) \text{ and } (\forall c' \in M(X))c \subseteq c' \Rightarrow c = c'\}.$$

Then define a controller f_n^* as

$$f_n^*(X) := \begin{cases} N(X) & X \in P \\ \Sigma_u & \text{otherwise.} \end{cases} \quad (17)$$

Notice that $f_n^*(X)$ does not contain any redundant control patterns.

Assume that for given state X , predicate P on \mathcal{X} and $e, e' \subseteq \Sigma_c$, checking $\delta(e, X) \in P, \delta(e, X)!$, $e \subseteq e'$, and $e = e'$ are basic computational steps. Then the complexities of computing both $M(X)$ and $N(X)$ are $O(2^{|\Sigma_c|} \times 2^{|\Sigma_c|})$. (Note that $|M(X)| \leq 2^{|\Sigma_c|}$.) The computational complexity of f_n^* is thus $O(|P| \times 2^{2|\Sigma_c|})$.

Theorem 9: The NDC f_n^* defined in (17) is the supremal element in $F(P)$.

Proof: We first show that $f_n^* \in F^+(P) \cap F^-(P)$. That $f_n^* \in F^+(P)$ follows from the fact that for any $X \in P$

$$(\forall \gamma \in f_n^o(X))(\exists \gamma' \in f_n^*(X))\gamma \subseteq \gamma'$$

$$f_n^o(X) \leq f_n^*(X).$$

That $f_n^* \in F^-(P)$ follows directly from the definition and Claim 3. It is also easy to see from the definition that f_n^* is balanced. Thus $f_n^* \in F(P)$.

Let $f_n \in F(P)$. Then $f_n \in F^-(P)$, and

$$(\forall X \in P)(\forall \gamma \in f_n(X))(\forall c \in 2^{\gamma})\delta(c, X) \in P \text{ or not } \delta(c, X)!.$$

From Claim 3

$$(\forall X \in P)(\forall \gamma \in f_n(X))(\forall c \in 2^{\gamma \cap \Sigma_c})\delta(c, X) \in P \text{ or not } \delta(c, X)!.$$

Therefore

$$(\forall X \in P)(\forall \gamma \in f_n(X))(\exists \gamma' \in N(X))\gamma \subseteq \gamma'$$

and $f_n \leq f_n^* (rel P)$. It follows that f_n^* is the supremal element in $F(P)$. \square

Therefore, for a nonempty controllable predicate P , $f_{n,con}^*$ will synthesize P and, among all controllers doing so, it will allow the plant G_{con} to have maximally concurrent closed-loop behavior.

For the ATM example, $\Sigma_c = \{\alpha_1, \alpha_2, \alpha_3\}$, and

$$M(X) = \begin{cases} \{\emptyset, \{\alpha_1\}, \{\alpha_2\}, \{\alpha_3\}, \{\alpha_1, \alpha_2\}, \\ \{\alpha_1, \alpha_3\}, \{\alpha_2, \alpha_3\}, \{\alpha_1, \alpha_2, \alpha_3\}\} & X \leq 2 \\ \{\emptyset, \{\alpha_1\}, \{\alpha_2\}, \{\alpha_3\}, \{\alpha_1, \alpha_2\}, \\ \{\alpha_1, \alpha_3\}, \{\alpha_2, \alpha_3\}\} & X = 3 \\ \{\emptyset, \{\alpha_1\}, \{\alpha_2\}, \{\alpha_3\}\} & X = 4 \\ \{\emptyset\} & X = 5 \end{cases}$$

$$f_n^*(X) := N(X) = \begin{cases} \{\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}\} & X \leq 2 \\ \{\{\alpha_1, \alpha_2, \alpha_4\}, \{\alpha_1, \alpha_3, \alpha_4\}, \\ \{\alpha_2, \alpha_3, \alpha_4\}\} & X = 3 \\ \{\{\alpha_1, \alpha_4\}, \{\alpha_2, \alpha_4\}, \\ \{\alpha_3, \alpha_4\}\} & X = 4 \\ \{\{\alpha_4\}\} & X = 5 \end{cases}$$

$$f_n^*(X) := \{\{\alpha_4\}\}, \quad X > 5$$

The control actions of $f_{n,con}^*$ are: for $X \leq 2$, enables all three RN's for transmission; at $X = 3$, randomly enables two RN's; at $X = 4$, randomly enables one RN; and for $X \geq 5$, disables all RN's. This is intuitively the best that any controller can possibly achieve in the presence of concurrency.

VI. CONCLUSIONS

The VDES is a compact model for a useful class of discrete-event control systems and allows computationally-efficient control synthesis [13], [14]. Continuing the study of VDES, this paper explored another feature of VDES: concurrency, or possible simultaneous occurrence of events. We formalized concurrency by extending the event set from Σ to $\Sigma_{con} := 2^\Sigma$; an extended event $e = \langle \alpha_1, \alpha_2, \dots, \alpha_k \rangle$ in Σ_{con} represents the simultaneous occurrence of $\alpha_1, \alpha_2, \dots, \alpha_k$. The VDES model incorporating concurrency is called concurrent VDES (CVDES).

It was shown that, owing to the effect of concurrency on control, a nonempty predicate on the state space can be synthesized by a (balanced) deterministic controller in a CVDES if and only if it is controllable and satisfies an additional condition called concurrent well-posedness. Since concurrently well-posed predicates are not closed under disjunction, it is not always possible to synthesize an optimal deterministic controller for a CVDES.

This led us to nondeterministic control of CVDES. It was shown that controllability is a necessary and sufficient condition for a predicate to be synthesized by a balanced nondeterministic controller in a CVDES. Therefore, optimal nondeterministic-controller synthesis (in terms of closed-loop predicate) is always achievable. Next we characterized the set of nondeterministic controllers synthesizing a given controllable predicate P , denoted by $F(P)$. A partial order was defined on this set to compare degrees of concurrency of the closed-loop system allowed by nondeterministic controllers. Finally, we proposed a formula for constructing the supremal element of $F(P)$. If P is a given legal specification, then the supremal element in $F(\text{sup}C(P))$ is a nondeterministic controller enforcing P which is optimal not only in the sense of reachable states but also in the sense of concurrency permitted to the controlled CVDES.

We remark that the results in this paper can be easily adapted to any automaton DES for which Claims 1, 2, and 3 can be established.

In future research it may be worthwhile to exploit the specific structure of CVDES so that the supremal element of $F(\text{sup}C(P))$ can be constructed in a computationally-efficient way, along the lines of [14].

Throughout this paper, we used a simple ATM example to demonstrate concepts and results. Though a small part of a complex system, this example does show that concurrency can arise naturally from and pose unique problems for the control of real-world systems such as telecommunication networks. For future research, it will be interesting to model larger systems using CVDES and address control specifications of wider scope.

REFERENCES

- [1] J. W. de Bakke, W. P. de Roever, and G. Rozenberg, Eds., *Current Trends in Concurrency* (Lecture Notes in Computer Science, Vol. 224). New York: Springer-Verlag, 1986.
- [2] C. H. Golaszewski and P. J. Ramadge, "Control of discrete event processes with forced events," in *Proc. 26th IEEE Conf. Decision and Control*, Los Angeles, CA, Dec. 1987, pp. 247-251.
- [3] L. E. Holloway and B. H. Krogh, "On closed-loop liveness of discrete event systems under maximally permissive control," in *Proc. 28th IEEE Conf. Decision and Control*, Tampa, FL, Dec. 1989, pp. 2725-2730.
- [4] R. Karp and R. Miller, "Parallel program schemata," *J. Comput. Syst. Sci.*, vol. 3, no. 4, pp. 167-195, 1969.
- [5] B. H. Krogh, "Controlled Petri nets and maximally permissive feedback logic," in *Proc. 25th Annu. Allerton Conf.*, Univ. Illinois at Urbana-Champaign, Oct. 1987, pp. 317-326.
- [6] B. H. Krogh and L. E. Holloway, "Synthesis of feedback control logic for discrete manufacturing systems," *Automatica*, vol. 27, no. 4, pp. 641-651, 1991.
- [7] S. LaFortune, "Modeling and analysis of transaction execution in database systems," *IEEE Trans. Automat. Contr.*, vol. 33, no. 5, pp. 439-447, 1988.
- [8] Y. Li, "On real-time supervisory control of discrete-event systems," M.A.Sc. thesis, Dep. Elec. Eng., Univ. Toronto, June 1986.
- [9] —, "Control of vector discrete-event systems," Ph.D. dissertation, Dep. Elec. Eng., Univ. Toronto, May 1991; available as Tech. Rep. 9106 (revised), Syst. Control Group, Dep. Elec. Eng., Univ. Toronto.
- [10] Y. Li and W. M. Wonham, "On real-time supervisory control of discrete-event systems," in *Proc. 1987 Amer. Control Conf.*, June 1987, pp. 1715-1720.
- [11] —, "On real-time supervisory control of discrete-event systems," *Inform. Sci.*, vol. 46, no. 3, pp. 159-183, 1988.
- [12] —, "Strict concurrency and nondeterministic control of discrete-event systems," in *Proc. 28th IEEE Conf. Decision and Control*, Tampa, FL, Dec. 1989, pp. 2731-2736.
- [13] —, "Control of vector discrete-event systems—Part I: The base model," *IEEE Trans. Automat. Contr.*, vol. 38, no. 8, pp. 1214-1227, Aug. 1993.
- [14] —, "Control of vector discrete-event systems—Part II: Controller synthesis," *IEEE Trans. Automat. Contr.*, vol. 39, no. 3, pp. 512-531, Mar. 1994.
- [15] A. Pattavina, "Nonblocking architectures for ATM switching," *IEEE Commun. Mag.*, vol. 31, no. 2, pp. 38-48, Feb. 1993.
- [16] P. J. Ramadge and W. M. Wonham, "Modular feedback logic for discrete event systems," *SIAM J. Contr. Optimization*, vol. 25, no. 5, pp. 1202-1218, 1987.
- [17] —, "The control of discrete event systems," *Proc. IEEE (Special Issue on Discrete Event Dynamic Systems)*, vol. 77, no. 1, pp. 81-98, Jan. 1989.
- [18] B. Selic, G. Gullekson, J. McGee, and I. Engelberg, "ROOM: An object-oriented methodology for developing real-time systems," in *Proc. CASE'92 Fifth Int. Workshop Computer-Aided Software Engineering*, Montreal, Quebec, July 1992.
- [19] T. Ushio and R. Matsumoto, "State feedback and modular control synthesis in controlled Petri nets," in *Proc. 27th IEEE Conf. Decision and Control*, Austin, TX, Dec. 1988, pp. 1502-1507.
- [20] T. Ushio, Y. Li, and W. M. Wonham, "Concurrency and state feedback in discrete-event systems," *IEEE Trans. Automat. Contr.*, vol. 37, no. 8, pp. 1180-1184, Aug. 1992.



Yong Li (S'86-M'90) was born in Sichuan, China, in 1962. He received the B.Eng. degree in computer science and automation from Chongqing University, Sichuan, in 1982, and the M.A.Sc. and Ph.D. degrees from the University of Toronto, Ontario, Canada, in 1986 and 1991, respectively, both in electrical engineering.

He joined Northern Telecom, Brampton, Ontario, in 1990, and has been with Bell-Northern Research, Ottawa, Ontario, since 1994. His current research interests include control of discrete-event systems,

software engineering for large-scale real-time systems, and telecommunications.

Dr. Li was a recipient of the Connaught Scholarship from the University of Toronto.



W. M. Wonham (M'64-SM'76-F'77) received the B.S. degree in engineering physics from McGill University, Montreal, P.Q., Canada, in 1956 and the Ph.D. degree in control engineering from the University of Cambridge, Cambridge, England, in 1961.

From 1961 to 1969, he was associated with the Control and Information Systems Laboratory at Purdue University, West Lafayette, IN, the Research Institute for Advanced Studies (RIAS) of the Martin Marietta Co., the Division of Applied Mathematics at Brown University, Providence, RI, and (as a National Academy of Sciences Research Fellow) with the Office of Control Theory and Application of NASA's Electronics Research Center. In 1970, he joined the Systems Control Group of the Department of Electrical Engineering at the University of Toronto, Ontario, Canada. He currently holds the J. Roy Cockburn Chair. In addition, he has held visiting academic appointments with the Department of Electrical Engineering at Massachusetts Institute of Technology, Cambridge; the Department of Systems Science and Mathematics at Washington University, St. Louis, MO; the Department of Mathematics of the University of Bremen; the Mathematics Institute of the Academia Sinica, Beijing; the Indian Institute of Technology, Kanpur; and other institutions. His research interests have lain in the areas of stochastic control and filtering, the geometric theory of linear multivariable control, and more recently in discrete event systems from the viewpoint of formal logic and language. He has authored or coauthored about 60 research papers as well as the book *Linear Multivariable Control: A Geometric Approach*.

Dr. Wonham is a Fellow of the Royal Society of Canada. In 1987, he was the recipient of the IEEE Control Systems Science and Engineering Award, and in 1990 was a Brouwer Medalist of the Netherlands Mathematical Society.

Structure of Model Uncertainty for a Weakly Corrupted Plant

Tong Zhou and Hidenori Kimura, *Fellow, IEEE*

Abstract—In this paper, we investigate the structure of the transfer function set which includes all eliminate the transfer functions deduced from the plant available information. It is shown that when an upper bound of the plant transfer function's H^∞ -norm has been supplied, and the noise contaminating the time domain identification experiment data is not too significant, such a transfer function set can be parameterized by a linear fractional transformation of two transfer function matrices. One of them is a fixed transfer function matrix which is completely determined by the plant available information and the noise magnitude. The other is a norm bounded, structure fixed, free transfer function matrix. Moreover, it is shown that the problem of analytically obtaining the fixed complexity nominal model that best approximates this transfer function set is as difficult as the μ -synthesis problem.

I. INTRODUCTION

ROBUSTNESS is one of the principal properties expected for control system, due to the high complexity of plant, plant parameter variation according to operation conditions, etc. In order to carry out such a study quantitatively, an initial step is to develop a mechanism to measure the size of the uncertainty of the plant nominal model error. Afterwards, it is necessary to investigate the problem of designing controllers for transfer function set defined by this mechanism and the problem of identifying the transfer function set measured by this mechanism from plant physically obtainable information.

In the last decade, most of the research effort has been focused on the problem of developing controller design theory and uncertainty measure mechanism. For example, gap metric, graph metric, H^∞ -norm, L^1 -norm, etc., have been developed to measure the uncertainty of nominal model error, and parametric uncertainty, additive unstructured uncertainty, multiplicative unstructured uncertainty, coprime factor unstructured uncertainty, etc., have been exploited to describe the model of a control plant with a nominal model, while H^∞ optimization theory, structured singular value theory (or μ), quadratic stability theory, etc., have been developed for robust controller design [4], [5], [8], [16], [18]. However, it is until recent years that the problem of identifying transfer function set measured by a mechanism suitable for robust

controller design from plant available information has begun to receive a considerable amount of attention, e.g., [3], [9], [12], [14], [15], [19], [21], [25], [26], [30], [32], [34], and [37]–[39].

Several approaches have been proposed to attack this problem. Briefly, these approaches can be divided into two kinds. One of them is a stochastic approach in which the uncertainty bound of the nominal model error is sometimes called “soft bound” [12], [24]. The other is a deterministic approach, and in this case, the nominal model error uncertainty bound is called “hard bound” [14], [15], [32], [21], [38], [39].

In a deterministic approach, there are two kinds of philosophy in identifying the transfer function set of a plant from its available information. One of them is to obtain the smallest transfer function set which can include all of the transfer functions that are deduced from the available information about the plant. This philosophy has a strong information theory support, because it means that with the increment of the information about the plant, the uncertainty about the plant dynamics will decrease [31], [35]. Another one is to find the smallest transfer function set that can not be falsified by the plant available information. This philosophy is based on a scientific principle which says that a model is correct if it can explain the observed plant input–output characteristics, and among the models that can not be falsified by the knowledge about the plant, the best choice is the “simplest” or “most powerful” model [25], [33]. The former is called worst case deterministic approach and the concepts are first proposed by Helmicki *et al.* in the field of identification for robust control [15]. The latter has been attacked by Zhou and Kimura. In that case, the problem of identifying the smallest unfalsified transfer function set has been reduced to a convex optimization problem, which is computationally tractable, under the condition that the denominator of the plant nominal model has been prescribed [37]–[39].

While it is generally impossible to obtain the “actual” model of a plant even in the ideal case, i.e., when a plant is linear and time invariant, because perfect information about a plant is usually not obtainable from measurement, due to the finiteness of identification experiment time length and unavoidable noise existent in identification experiment data, both the transfer function sets, which are deduced by the philosophy of unfalsified modeling and the philosophy of worst case deterministic approach, respectively, will play important roles in robust controller design. This is based on

Manuscript received August 20, 1993; revised March 2, 1994 and August 12, 1994. Paper recommended by Associate Editor, A. L. Tits.

T. Zhou is with the Department of Automation, University of Electronic Science and Technology of China, Chengdu, Sichuan Province 610054, People's Republic of China.

H. Kimura is with the Department Mathematical Engineering and Information Physics, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113, Japan.

IEEE Log Number 9408277.

the following arguments. That is, the worst case deterministic approach will give a transfer function set that can guarantee the actual plant performances when a controller is designed such that the desirable performances are met for all of the transfer functions belonging to the transfer function set, provided that the actual plant dynamics can be appropriately approximated by a transfer function. On the other hand, the unfalsified modelling will bring out a transfer function set that a controller should be designed to make every transfer function in that transfer function set satisfy the desirable performances in order to let the actual plant achieve these performances.

In unfalsified modelling, the well known Carathéodory–Fejér extrapolation theorem (or Schur–Takagi–AAK Theorem) plays an essential role [13], [22], [2]. The importance of this theorem in identification and model validation has been independently found by Poolla *et al.* and Zhou *et al.* [28], [36], [37].

It is still a challenging problem, however, to obtain the smallest transfer function set which can include all eliminate the transfer functions that are deduced from the available information of a plant and is compatible with robust controller design, although a number of robustly convergent algorithms have been proposed, e.g., [14], [15]. Basically, such a problem is to approximate a function set by a single function belonging to a prescribed set, e.g., this function should be in \mathcal{H}^∞ , and/or it should have a complexity as low as possible for the convenience of robust controller design, etc. Therefore, the structure of the transfer function set will play essential roles in finding out the identification algorithm. For instance, in information-based complexity theory, it is well known that if a function set is symmetric, then it will be best approximated by its center, no matter what metric is used in measuring the approximation error, provided that there is no restriction imposed on the function that is used to approximate the function set [31].

In this paper, we investigate the problem concerning the structure of the transfer function set which includes all of the possible transfer functions of a plant, based on some kinds of the plant obtainable information. It is shown that when an upper bound of the \mathcal{H}^∞ -norm of the plant transfer function has been given and the plant time-domain input–output has been measured over a finite time interval, if it is only the plant output that has been contaminated by a not too significant magnitude bounded noise, then this transfer function set can be described by a linear fractional transformation of two transfer function matrices, one of them is completely determined by the plant available information and the noise magnitude and is therefore fixed, another one is uncertain but norm bounded and structure fixed. Moreover, it is also shown that the problem of obtaining the fixed complexity plant nominal model that best approximates this transfer function set has an equal level of difficulty as that met in μ -synthesis problem.

An outline of this paper is as follows. In Section II, the problem is formulated. The structure of the transfer function set is investigated in Section III, while the application of the main results to identification for robust control is discussed in Section IV. Finally, conclusions of this paper are given in Section V.

Many of the results of Section III of this paper have appeared, without proof, in [40].

Notations

\mathcal{R}	The set of real numbers.
\mathcal{C}	The set of complex numbers.
$\mathcal{R}^{m \times n}$	The set of $m \times n$ real matrices.
$\mathcal{C}^{m \times n}$	The set of $m \times n$ complex matrices.
\mathcal{D}	The closed unit disk defined as $\mathcal{D} = \{z \mid z \leq 1\}$.
I_k	Identity matrix of $k \times k$ dimension. In cases in which confusion will not result, the subscript k is omitted.
	Scalar 0 or matrix with suitable dimension and all elements being zero.
$\text{diag}\{\alpha_1, \alpha_2, \dots, \alpha_n\}$	Diagonal matrix or block diagonal matrix defined by scalars or matrices $\alpha_1, \alpha_2, \dots, \alpha_n$.
X^t	Transpose of vector or matrix X .
$\det(X)$	The determinant of a square matrix X .
$\bar{\sigma}(X)$	The maximal singular value of matrix X .
\mathcal{H}^∞	Transfer function set or transfer function matrix set formed by transfer function or transfer function matrix that is analytic on the closed unit disk $\mathcal{D} = \{z \mid z \leq 1\}$.
$\ F\ $	The \mathcal{H}^∞ -norm of transfer function $F(z) \in \mathcal{H}^\infty$ which is defined as

$$\|F(z)\|_\infty = \max_{|z| \leq 1} |F(z)|.$$

\mathcal{BH}^∞	Transfer function set which consists of all transfer functions that are analytic when $ z \leq 1$ and have \mathcal{H}^∞ -norm not greater than 1, i.e., $\mathcal{BH}^\infty = \{g(z) \mid \ g(z)\ _\infty \leq 1\}$.
-----------------------	---

Linear fractional transformation (LFT)

$$\mathcal{F}_1 \left(\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}, K \right) = P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21}.$$

Homographic transformation (HM)

$$\mathcal{HM} \left(\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}, K \right) = (P_{11}K + P_{12})(P_{21}K + P_{22})^{-1}.$$

Structured singular value ($\mu_N(M)$): For given positive integers $c_1, c_2, \dots, c_T; r_1, r_2, \dots, r_S; m_1, m_2, \dots, m_F$, define a matrix structure

$$\begin{aligned} \mathcal{X} = \{ \Delta \mid \Delta = & \text{diag}\{\delta_1 I_{c_1}, \dots, \delta_S I_{c_T}, \gamma_1 I_{r_1}, \dots, \\ & \gamma_S I_{r_S}, \Delta_1, \dots, \Delta_F\}, \\ & \delta_i \in \mathcal{R}, \gamma_j \in \mathcal{C}, \Delta_k \in \mathcal{C}^{m_k \times m_k}, \\ & 1 \leq i \leq T, 1 \leq j \leq S, 1 \leq k \leq F \}. \end{aligned}$$

Then, the structured singular value of matrix M with dimension

$$\left(\sum c_i + \sum r_i + \sum m_i \right) \times \left(\sum c_i + \sum r_i + \sum_{i=1}^F m_i \right)$$

is defined as

$$\mu_{\lambda}(M) := \begin{cases} \frac{1}{\min \{ \bar{\sigma}(\Delta), \Delta \in \mathcal{X}, \det(1 - M\Delta) = 0 \}}, \\ 0, & \text{if there is no } \Delta \in \mathcal{X} \text{ which makes} \\ & I - M\Delta \text{ singular.} \end{cases}$$

II. PROBLEM FORMULATION

Mathematically, the problem of identifying a nominal model and its error uncertainty bound in the worst case by a deterministic approach is a problem to best approximate a function set by a suitable single function under a metric compatible with robust controller design. While the transfer function that best approximates the prescribed transfer function set gives the plant nominal model, the approximation error will provide the uncertainty bound of the plant nominal model error. Both of them are essential in robust controller design. On the other hand, some restrictions are generally imposed on the function that is used to approximate the function set, which is determined by this function's intended application. For example, in robust controller design, a lower complexity nominal model is commonly highly appreciated; moreover, it is usually required to let the plant nominal model have the same number of unstable poles as that of the plant "actual" transfer function; and so on. In these senses, such an identification problem differs significantly from the traditional identification problem, in which it is desirable to find some parameters that minimizes the "prediction error" of the nominal model which is defined by *a priori* chosen model structure and the parameters to be found [24], [7].

To approximate a function set, it is preferable to investigate the structure of the function set first, because the structure of the function set plays an essential role in developing the approximation algorithm. For example, a symmetric function set will always be best approximated by its center, no matter what metric is utilized to measure the approximation error, and it is generally hard to find out the optimal approximation algorithm for a function set with arbitrary structure [31]. In worst-case deterministic identification, the function set is a transfer function set and every transfer function belonging to it is deduced from the plant physically obtainable information.

Generally, two kinds of information can be obtained from a plant. One of them is in frequency domain and another in time domain.

Plant frequency domain information can be obtained by applying sinusoidal input with frequency corresponding to the frequency point of interest.

The problem that will be investigated here is formulated as follows.

Assume that the plant is linear time invariant and stable. Moreover, assume that the \mathcal{H}^{∞} -norm of the plant transfer function is not greater than γ . Given a series of time domain identification experiment data of the plant

$$(u_0, y_0), (u_1, y_1), \dots, (u_n, y_n), \quad u_0 \neq 0$$

which is corrupted by an additive magnitude bounded noise series

$$(v_0, v_1, v_2, \dots, v_n), \quad v_i \in \mathcal{R}, |v_i| \leq \kappa, 0 \leq i \leq n$$

in the plant output y_i , $0 \leq i \leq n$. The problem is to find out all of the transfer functions that do not contradict the aforementioned information about the plant.

From the problem formulation, it is obvious that the transfer function set that consists of all the plant possible transfer functions can be approximated, which means that a single transfer function can be found that approximates the transfer function set with a finite error. For instance, the zero transfer function will approximate this transfer function set with an error not greater than γ , provided that the error is measured by \mathcal{H}^{∞} -norm. Moreover, when the plant perfect information is supplied, that is, when $\kappa \rightarrow 0$ and $n \rightarrow \infty$, the transfer function set will degenerate into one transfer function. In fact, this transfer function is the plant actual transfer function. Furthermore, compared with the maximal magnitude of the plant unit impulse response M and its relative stability margin ρ which are widely used as the *a priori* information about the plant in the literature dealing with worst-case deterministic identification problem, e.g., [14] and [15], an upper bound of the \mathcal{H}^{∞} -norm of the plant transfer function is easier to obtain. Note that when a plant is linear time invariant and stable, according to the well known maximum-modulus theorem [29], an upper bound of the plant transfer function's \mathcal{H}^{∞} -norm can be obtained by measuring the plant frequency response, which is generally possible. In these senses, the formulated problem is well-posed.

III. STRUCTURE OF THE TRANSFER FUNCTION SET

Under the condition that the plant is linear and time invariant during the identification experiment, the uncertainty about the plant transfer function results from two factors. First, the measure time length of identification experiment is limited, which means that n is generally a finite integer. Therefore, we can only obtain finite sampling points of the plant unit impulse response. The plant remainder unit impulse response will make us uncertain about its transfer function. Second, there always exists effect of noise during identification experiment, which implies that we can only obtain the intervals in which the first n sampling points of the actual plant unit impulse response exist, and we are also uncertain about the real value of the first n sampling points of the plant unit impulse response.

As the first step, we shall investigate the uncertainty on the plant transfer function caused by the finite length of identification experiment time length.

This problem is closely connected to a classical extrapolation problem, Schur extrapolation problem [2], [6], [17]. First, we introduce the following results which are obtained by Adamjan *et al.* [1] and [22].

AAK Theorem: For given real number g_0, g_1, \dots, g_n , define

$$\bar{G} = \begin{bmatrix} g_n & g_{n-1} & g_1 & g_0 \\ g_{n-1} & g_{n-2} & g_0 & 0 \\ \vdots & \vdots & & \\ g_1 & g_0 & 0 & \\ g_0 & 0 & 0 & \end{bmatrix}$$

If $\bar{\sigma}(\bar{G}) < \gamma$, then

$$g(z) = \gamma z^{n+1} \frac{p(\frac{1}{z})\epsilon(z) + q(\frac{1}{z})}{q(z)\epsilon(z) + p(z)}, \quad \epsilon(z) \in \mathcal{BH}^\infty \quad (1)$$

exhausts all the transfer functions

$$g(z) = g_0 + g_1 z + g_2 z^2 + \cdots + g_n z^n + z^{n+1} \sum_{j=1}^{\infty} h_j z^j, \quad (2)$$

$$\leq \gamma. \quad (3)$$

Moreover, every transfer function which has the form expressed in (1) satisfies the conditions represented in (2) and (3). Here

$$p(z) = \sum_{j=1}^{\infty} p_j z^{j-1}, \quad q(z) = \sum_{j=1}^{\infty} q_j z^{j-1},$$

$$p = [p_1 \ p_2 \ \cdots]^t = \gamma R e, \quad q = [q_1 \ q_2 \ \cdots]^t = T G R e.$$

$$R = (\gamma^2 I - G^2)^{-1}, \quad e = [1 \ 0 \ 0 \ \cdots]^t,$$

$$G = \begin{bmatrix} g_n & g_{n-1} & \cdots & g_1 & g_0 & 0 & \cdots \\ g_{n-1} & g_{n-2} & \cdots & g_0 & 0 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ g_1 & g_0 & \cdots & 0 & 0 & 0 & \cdots \\ g_0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$T = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

In retrospect, the aforementioned extrapolation problem is an extension of the Carathéodory–Fejér extrapolation problem and has been well investigated by Schur *et al.* Moreover, it is well known as a classical moment problem related to the Nevanlinna–Pick interpolation problem [1], [2], [13], [22]. This classical moment problem has found wide applications in broad-band matching, cascade synthesis, network and semiconductor modeling, etc. [17]. The ideas and concepts developed in solving these classical moment problems have accepted a wide attention in control community, also. For example, the concepts of Hankel operator, Toeplitz operator, J-lossless factorization, J-spectral factorization, etc., have been widely used in solving the so-called \mathcal{H}^∞ optimization problem, robust stabilization problem, model reduction problem, etc. [5], [10], [11], [16], [20], [23]. The solution of this extrapolation problem was first given by Adamjan *et al.* in the explicit form expressed in (1) [1], [22], and a proof based on elementary matrix manipulation can be found in [41].

Based on the AAK Theorem, we investigate the uncertainty on the plant transfer function caused by the unmeasured plant output.

Theorem 1: Assume that $Y_0^t Y_0 < \gamma^2 U^t U$. Then, every transfer function which is consistent with the given input–output data $(u_i, y_i^0) \mid_{i=0}^n$, $u_0 \neq 0$, and has \mathcal{H}^∞ -norm not greater than γ can be represented by

$$g(z) = \gamma \frac{\sum_{j=1}^{n+1} \hat{q}_j^0 z^{j-1} + (\sum_{j=1}^{n+1} \hat{p}_j^0 z^{n+2-j}) \epsilon(z)}{\sum_{j=1}^{n+1} \hat{p}_j^0 z^{j-1} + (\sum_{j=1}^{n+1} \hat{q}_j^0 z^{n+2-j}) \epsilon(z)}, \quad \epsilon(z) \in \mathcal{BH}^\infty.$$

Moreover, every transfer function belonging to transfer function set

$$\begin{aligned} \mathcal{G}_n(\gamma) &= \{g(z) \mid g(z) \\ &= \gamma \frac{\sum_{j=1}^{n+1} \hat{q}_j^0 z^{j-1} + (\sum_{j=1}^{n+1} \hat{p}_j^0 z^{n+2-j}) \epsilon(z)}{\sum_{j=1}^{n+1} \hat{p}_j^0 z^{j-1} + (\sum_{j=1}^{n+1} \hat{q}_j^0 z^{n+2-j}) \epsilon(z)}, \\ &\quad \epsilon(z) \in \mathcal{BH}^{mfly}\} \end{aligned}$$

has \mathcal{H}^∞ -norm not greater than γ and matches the given input–output data. Here

$$p^0 = [\hat{p}_1^0 \ \hat{p}_2^0 \ \cdots \ \hat{p}_{n+1}^0]^t = \gamma U (\gamma^2 U^t U - Y_0^t Y_0)^{-1} e,$$

$$\hat{q}^0 = [\hat{q}_1^0 \ \hat{q}_2^0 \ \cdots \ \hat{q}_{n+1}^0]^t = Y_0 (\gamma^2 U^t U - Y_0^t Y_0)^{-1} e,$$

$$U = \begin{bmatrix} u_0 & 0 \\ u_1 & u_0 \\ \vdots & \vdots \\ u_n & u_{n-1} \end{bmatrix}$$

$$Y_0 = \begin{bmatrix} y_n^0 & y_{n-1}^0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & \vdots \\ 0 & 0 \end{bmatrix}.$$

Proof: For a plant with transfer function $G(z) = \sum_{j=0}^{\infty} g_j z^j$, if $(u_i, y_i^0) \mid_{i=0}^n$ are the input–output pairs of the plant, then, it is necessary that

$$\begin{bmatrix} y_0^0 & g_0 & 0 & \cdots & u_0 \\ y_1^0 & g_1 & g_0 & \cdots & u_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_n^0 & g_n & g_{n-1} & \cdots & u_n \end{bmatrix} \quad (4)$$

that is

$$\begin{bmatrix} y_0^0 & y_0^0 \\ y_1^0 & y_0^0 \\ \vdots & \vdots \\ y_n^0 & y_{n-1}^0 \end{bmatrix} = \begin{bmatrix} g_0 & 0 & \cdots & 0 \\ g_1 & g_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g_n & g_{n-1} & \cdots & g_0 \end{bmatrix} \begin{bmatrix} u_0 & 0 & \cdots & 0 \\ u_1 & u_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_n & u_{n-1} & \cdots & u_0 \end{bmatrix} \quad (5)$$

Since $u_0 \neq 0$, we have

$$\hat{G} = Y_0 U^{-1} = U^{-1} Y_0. \quad (6)$$

Here

$$\hat{G} = \begin{bmatrix} g_0 & 0 & \cdots & 0 \\ g_1 & g_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g_n & g_{n-1} & \cdots & g_0 \end{bmatrix}.$$

On the other hand, it can also be proved that if

$$\hat{G} = Y_0 U^{-1} = U^{-1} Y_0, \quad (7)$$

then, when $u_i |_{i=0}^n$ are applied to the plant, the output of the plant will be $y_i^0 |_{i=0}^n$.

Therefore, a plant with transfer function $g(z) = \sum_{j=0}^{\infty} g_j z^j$ has input-output pairs $(u_i, y_i^0) |_{i=0}^n$, $u_0 \neq 0$, if and only if

$$\hat{G} = Y_0 U^{-1} = U^{-1} Y_0. \quad (8)$$

Note that

$$\begin{bmatrix} g_n & \cdots & g_1 & g_0 \\ g_{n-1} & \cdots & g_0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ g_0 & \cdots & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} g_0 & 0 & \cdots & 0 \\ g_1 & g_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g_n & g_{n-1} & \cdots & g_0 \end{bmatrix}. \quad (9)$$

Let

$$T_1 = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{bmatrix}$$

then using the same notations as those in the AAK Theorem, we have

$$\bar{G} = T_1 G = T_1 Y_0 U^{-1} = T_1 U^{-1} Y_0. \quad (10)$$

From

$$G = \begin{bmatrix} \bar{G} & 0 \\ 0 & 0 \end{bmatrix}$$

we conclude

$$G^2 = \begin{bmatrix} \bar{G} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{G} & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \bar{G}^2 & 0 \\ 0 & 0 \end{bmatrix}. \quad (11)$$

$$R = (\gamma^2 I - G^2)^{-1} = \begin{bmatrix} (\gamma^2 I_{n+1} - \bar{G}^2)^{-1} & 0 \\ 0 & \gamma^{-2} I_{\infty} \end{bmatrix}, \quad (12)$$

$$\begin{aligned} p &= \gamma R e = \begin{bmatrix} \gamma(\gamma^2 I_{n+1} - \bar{G}^2)^{-1} & 0 \\ 0 & \gamma^{-1} I_{\infty} \end{bmatrix} \begin{bmatrix} \hat{e} \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \gamma(\gamma^2 I_{n+1} - \bar{G}^2)^{-1} \hat{e} \\ 0 \end{bmatrix}, \end{aligned} \quad (13)$$

$$\begin{aligned} q &= T G R e = T \begin{bmatrix} \bar{G} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} (\gamma^2 I_{n+1} - \bar{G}^2)^{-1} & 0 \\ 0 & \gamma^{-2} I_{\infty} \end{bmatrix} \begin{bmatrix} \hat{e} \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \bar{G}(\gamma^2 I_{n+1} - \bar{G}^2)^{-1} \hat{e} \\ 0 \end{bmatrix}. \end{aligned} \quad (14)$$

Note that

$$\bar{G}^2 = \bar{G}^t \bar{G} = U^{-t} Y_0^t Y_0 U^{-1}. \quad (15)$$

Define

$$\bar{p} = \gamma(\gamma^2 I_{n+1} - \bar{G}^2)^{-1} \hat{e} = \gamma U(\gamma^2 U^t U - Y_0^t Y_0)^{-1} U^t \hat{e}, \quad (16)$$

$$\bar{q} = \bar{G}(\gamma^2 I_{n+1} - \bar{G}^2)^{-1} \hat{e} = T_1 Y_0(\gamma^2 U^t U - Y_0^t Y_0)^{-1} U^t \hat{e}. \quad (17)$$

On the other hand

$$U^t e = \begin{bmatrix} u_0 & 0 & \cdots & 0 \\ u_1 & u_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_n & u_{n-1} & \cdots & u_0 \end{bmatrix}^t \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = u_0 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = u_0 \hat{e} \quad (18)$$

which implies that

$$\bar{p} = u_0 p^0, \quad (19)$$

$$\bar{q} = u_0 T_1 \hat{q}^0. \quad (20)$$

Therefore

$$p(z) = \sum_1^{\infty} p_j z^{j-1} = u_0 \sum_1^{n+1} p_j^0 z^{j-1}, \quad (21)$$

$$p\left(\frac{1}{z}\right) = u_0 \sum_1^{n+1} p_j^0 z^{1-j}, \quad (22)$$

$$q(z) = \sum_1^{\infty} q_j z^{j-1} = u_0 \sum_1^{n+1} \hat{q}_j^0 z^{n+2-j}, \quad (23)$$

$$q\left(\frac{1}{z}\right) = u_0 \sum_1^{\infty} \hat{q}_j^0 z^{j-(n+2)}. \quad (24)$$

From the AAK Theorem, we conclude that every transfer function which is consistent with $(u_i, y_i^0) |_{i=0}^n$, $u_0 \neq 0$, and has \mathcal{H}^{∞} -norm not greater than γ can be represented as

$$\begin{aligned} g(z) &= \gamma z^{n+1} \frac{u_0 \sum_1^{n+1} \hat{q}_j^0 z^{-(n+2)} + (u_0 \sum_1^{n+1} \hat{p}_j^0 z^{1-j}) \epsilon(z)}{u_0 \sum_1^{n+1} \hat{p}_j^0 z^{j-1} + (u_0 \sum_1^{n+1} \hat{q}_j^0 z^{n+2-j}) \epsilon(z)} \\ &= \gamma \frac{\sum_1^{n+1} \hat{q}_j^0 z^{j-1} + (\sum_1^{n+1} \hat{p}_j^0 z^{n+2-j}) \epsilon(z)}{\sum_1^{n+1} \hat{p}_j^0 z^{j-1} + (\sum_1^{n+1} \hat{q}_j^0 z^{n+2-j}) \epsilon(z)}, \\ &\quad \epsilon(z) \in \mathcal{BH}^{\infty} \end{aligned} \quad (25)$$

and every transfer function belonging to transfer function set $\mathcal{G}_n(\gamma)$ has \mathcal{H}^{∞} -norm not greater than γ and is consistent with the given input-output data.

This completes the proof. \blacksquare

Theorem 1 implies that when there is no noise contaminating the identification data, then all of the transfer functions

that are consistent with the identification experiment data and have \mathcal{H}^∞ -norm not greater than γ can be expressed by a homographic transformation of two transfer function matrices, one of them is a fixed transfer function matrix which is completely determined by the upper bound of the plant transfer function's \mathcal{H}^∞ -norm and the identification experiment data, another one is a free transfer function except that its \mathcal{H}^∞ -norm is bounded by 1.

The results in Theorem 1 can be extended to the most general case, i.e., when $Y_0^t Y_0 \leq \gamma^2 U^t U$. According to the input-output extrapolation-minimization theorem [28], [36], [37] (in [28], it is called the extension theorem), in the case that $\bar{\sigma}(Y_0 U^{-1}) = \gamma$, there is only one transfer function that is consistent with the prescribed input-output pairs and has \mathcal{H}^∞ -norm not greater than γ , represent it as $f(z)$. In fact, in this case, the \mathcal{H}^∞ -norm of this transfer function is equal to γ . This fact implies that when $\bar{\sigma}(Y_0 U^{-1}) = \gamma$, we can completely determine the plant transfer function from its input-output data, provided that there is no noise existent in the identification experiment data (however, this is extremely rare in actual identification process). Note that $f(z) \in \mathcal{G}_n(\gamma + \alpha)$ for arbitrary $\alpha > 0$, and $\mathcal{G}_n(\gamma + \alpha_1) \subset \mathcal{G}_n(\gamma + \alpha_2)$ whenever $\alpha_1 < \alpha_2$ (this will be shown in the next corollary), it can be proved that $\{f(z)\} = \lim_{\alpha \rightarrow 0^+} \mathcal{G}_n(\gamma + \alpha) = \mathcal{G}_n(\gamma)$, which is to be shown. A detailed statement of this proof can be found in [41].

The condition $Y_0^t Y_0 \leq \gamma^2 U^t U$ for the transfer function set parameterization is natural. From the input-output extrapolation-minimization theorem [28], [36], [37], it can be claimed that if $Y_0^t Y_0 > \gamma^2 U^t U$, then no transfer function exists that matches the given input-output data and has \mathcal{H}^∞ -norm not greater than γ . Therefore, to prevent the transfer function set $\mathcal{G}_n(\gamma)$ from being void, it is necessary to make the assumption $Y_0^t Y_0 \leq \gamma^2 U^t U$.

Now, we would like to investigate some of the properties of the transfer function set $\mathcal{G}_n(\gamma)$ defined in Theorem 1. First, we discuss the relation between $\mathcal{G}_n(\gamma)$ and γ .

Corollary 1: Let matrices U , Y_0 and vector ϵ be the same as those defined in Theorem 1. Assume that $\gamma_1^2 U^t U > Y_0^t Y_0$, $\gamma_2^2 U^t U > Y_0^t Y_0$. Define transfer function sets $\mathcal{G}_n(\gamma_1)$, $\mathcal{G}_n(\gamma_2)$ through γ_1 , U , Y_0 ; γ_2 , U , Y_0 , respectively, as $\mathcal{G}_n(\gamma)$ in Theorem 1. If $\gamma_1 \geq \gamma_2$, then $\mathcal{G}_n(\gamma_1) \supseteq \mathcal{G}_n(\gamma_2)$.

Proof: Let

$$= U^{-1} \begin{bmatrix} y_0^0 & \gamma \\ y_1^0 \end{bmatrix} \quad (26)$$

From Theorem 1, it is obvious that

$$\mathcal{G}_n(\gamma_1) = \{g(z) \mid g(z) = g_0 + g_1 z + \cdots + g_n z^n + z^n \sum_{i=0}^{\infty} h_i z^i, \quad \|g(z)\|_\infty \leq \gamma_1\} \quad (27)$$

$$\mathcal{G}_n(\gamma_2) = \{g(z) \mid g(z) = g_0 + g_1 z + \cdots + g_n z^n + z^n \sum_{i=0}^{\infty} h_i z^i, \quad \|g(z)\|_\infty \leq \gamma_2\} \quad (28)$$

Since $\gamma_1 \geq \gamma_2$, we have

$$\mathcal{G}_n(\gamma_1) \supseteq \mathcal{G}_n(\gamma_2). \quad (29)$$

This completes the proof. ■

Corollary 1 tells us that when we apply sinusoidal input to a plant in order to obtain an upper bound of the \mathcal{H}^∞ -norm of the plant transfer function, and we are not sure about it, then, it is safer to choose a larger one, in order that the actual plant transfer function is included in the transfer function set that will be approximated.

The next corollary suggests that if there is no noise in the identification experiment data and the identification experiment data length intends to infinity, then the transfer function set $\mathcal{G}_n(\gamma)$ which is deduced from the plant available information will degenerate into one transfer function, i.e., the plant actual transfer function.

Corollary 2: For given input-output data $(u_i, y_i^0) \mid_{i=0}^{n1}$, $(u_i, y_i^0) \mid_{i=0}^{n2}$, and a positive number γ , define matrices U_{n1} , Y_{n10} ; U_{n2} , Y_{n20} ; transfer function sets $\mathcal{G}_{n1}(\gamma)$, $\mathcal{G}_{n2}(\gamma)$ as those in Theorem 1. If $u_0 \neq 0$, $Y_{n10}^t Y_{n10} < \gamma^2 U_{n1}^t U_{n1}$, $Y_{n20}^t Y_{n20} < \gamma^2 U_{n2}^t U_{n2}$, $n2 \leq n1$, then, $\mathcal{G}_{n1}(\gamma) \subseteq \mathcal{G}_{n2}(\gamma)$.

Proof: Let

$$\begin{aligned} & \begin{bmatrix} g_0 \\ g_1 \end{bmatrix} \\ & = U_{n2}^{-1} \begin{bmatrix} y_1^0 \\ \vdots \\ y_{n2}^0 \end{bmatrix} \end{aligned} \quad (30)$$

Since $n1 \geq n2$, from the structure of matrices U_{n1} , U_{n2} , it is obvious that

$$U_{n1} \begin{bmatrix} U_{n2} & 0 \\ X_1 & X_2 \end{bmatrix} \quad (31)$$

Here

$$\begin{aligned} X_1 &= \begin{bmatrix} u_{n2+1} & u_{n2} & \cdots & u_1 \\ u_{n2+2} & u_{n2+1} & \cdots & u_2 \end{bmatrix} \\ & \quad \begin{bmatrix} u_{n1} & u_{n1-1} & \cdots & u_{n1-n2} \end{bmatrix} \\ X_2 &= \begin{bmatrix} u_0 & 0 & \cdots & 0 \\ u_1 & u_0 & \cdots & 0 \end{bmatrix} \\ & \quad \begin{bmatrix} u_{n1-n2-1} & u_{n1-n2-2} & \cdots & u_0 \end{bmatrix} \end{aligned}$$

Therefore

$$U_{n1}^{-1} = \begin{bmatrix} U^{-1} & 0 \\ -X_2^{-1} X_1 U_{n2}^{-1} & X_2^{-1} \end{bmatrix} \quad (32)$$

Hence

$$g_i' = g_i, \quad 0 \leq i \leq n2. \quad (33)$$

From Theorem 1, we have

$$\mathcal{G}_{n1}(\gamma) = \left\{ q(z) \mid q(z) = q_0 + q_1 z + \dots + q_{n1} z^{n1} + z^{n1+1} \sum_{i=0}^{\infty} h_i z^i, \quad \|q(z)\|_{\infty} \leq \gamma \right\} \quad (34)$$

$$\mathcal{G}_{n2}(\gamma) = \left\{ q(z) \mid q(z) = q_0 + q_1 z + \dots + q_{n2} z^{n2} + z^{n2+1} \sum_{i=0}^{\infty} h_i z^i, \quad \|q(z)\|_{\infty} \leq \gamma \right\} \quad (35)$$

From $n1 \geq n2$, we conclude that

$$\mathcal{G}_{n1}(\gamma) \subseteq \mathcal{G}_{n2}(\gamma) \quad (36)$$

This completes the proof. ■

Corollary 2 also implies that as the information about the plant increasing, the "size" of the transfer function set that includes all of the possible plant transfer functions or the uncertainty about the plant transfer function will generally decrease. This is consistent with physical intuition.

From the above investigation, we see that all of the possible plant transfer functions are parameterized by a fixed homographic transformation with a norm bounded free transfer function and the transfer function set that consists of all the plant possible transfer functions will monotonically degenerate into the actual plant transfer function as the identification experiment time increasing to infinity in the case that the measured plant input/output pairs are not corrupted by noise.

Now, we investigate the variation of the homographic transformation in the face of noise that is existent in the measured plant output.

First we consider the change of vectors p^0, q^0 defined in Theorem 1.

The next well known fact plays an important role in developing the relation between the variation of vector p^0, q^0 and the noise in the measured plant output.

Fact 1 For matrices $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}$ if $(I_n + AB)^{-1}$ exists, then $B(I_m + 1B)^{-1} = (I_n + B1)^{-1}B$.

For brevity, we assume without loss of generality that $\gamma = 1$ ¹.

The variation of vectors p^0, q^0 according to the noise existent in the measured plant output is indicated in Theorem 2.

Theorem 2 Assume that $y_t^* = y_t + v_t, |v_t| \leq \kappa, 0 \leq t \leq n$. Moreover assume that $\max_{|t| < \kappa} \sigma(Y, U^{-1}) < 1$. Let

$$p^* = U(U^t U - Y_r^t Y_r)^{-1} e, \quad q^* = Y_r(U^t U - Y_r^t Y_r)^{-1} e$$

$$p = U(U^t U - Y^t Y)^{-1} e, \quad q = Y(U^t U - Y^t Y)^{-1} e$$

Then, for all possible vectors p^* and q^*

$$\begin{bmatrix} p^* \\ q^* \end{bmatrix} = \begin{bmatrix} p \\ q \end{bmatrix} - \begin{bmatrix} U^t - Y^t & \\ -Y & U \end{bmatrix}^{-1} W_1 \bar{\Lambda} \left[I_{2(n+1)} + W_1 \begin{bmatrix} U^t - Y^t & \\ -Y & U \end{bmatrix}^{-1} W_1 \bar{\Lambda} \right]^{-1} W_1 \begin{bmatrix} p \\ q \end{bmatrix}$$

¹When $\gamma \neq 1$ we can normalize the plant output and the noise magnitude in order to make $\gamma = 1$. In detail we may use $y_t/\gamma, \kappa/\gamma$ instead of y_t, κ respectively.

Here, $\bar{\Lambda}$ is a norm bounded uncertain matrix defined by

$$\bar{\Lambda} = \text{diag}\{\delta_0 I_{2(n+1)^2}, \delta_1 I_{2(n+1)^2}, \dots, \delta_n I_{2(n+1)^2}\}, \quad \delta_i \in \mathcal{R}, |\delta_i| \leq 1, 0 \leq i \leq n$$

and the other matrices are constant real matrices defined by $u_i, |i|=0, y_i, |i|=0, y_i^*, |i|=0, n$ and κ as

$$Y_t = \begin{bmatrix} y_0^* & 0 & 0 \\ y_1^* & y_0^* & 0 \\ \vdots & \vdots & \vdots \\ y_n^* & y_{n-1}^* & y_0^* \end{bmatrix}, \quad Y = \begin{bmatrix} y_0 & 0 & 0 \\ y_1 & y_0 & 0 \\ \vdots & \vdots & \vdots \\ y_n & y_{n-1} & y_0 \end{bmatrix},$$

$$W_t = \begin{bmatrix} 0 & E_t \\ F_t^t & 0 \end{bmatrix} \bar{T}_t, \quad W_r = \bar{T}_r \begin{bmatrix} E_r^t & 0 \\ 0 & E_r \end{bmatrix}$$

$$I_t = [I_{n+1} \quad I_2 \quad I_2^t] \quad E_t = \kappa [I_{n+1} \quad I_{n+1} \quad I_{n+1}]^t,$$

$$I_2 = \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{n+1} \Bigg\}^{n+1}$$

$$\bar{T}_t = [I_{2(n+1)} \quad I_{2(n+1)} \quad I_{2(n+1)}] \quad \bar{T}_r = [\bar{\gamma}_0 \quad \bar{\gamma}_1 \quad \bar{\gamma}_n]^t,$$

$$\bar{\gamma}_0 = \text{diag}\{I_{n+1} \quad 0 \quad 0 \quad I_{n+1} \quad 0 \quad 0\},$$

$$\bar{\gamma}_1 = \text{diag}\{0 \quad I_{n+1} \quad 0 \quad 0 \quad I_{n+1} \quad 0\},$$

$$\bar{\gamma}_n = \text{diag}\{0 \quad 0 \quad I_{n+1} \quad 0 \quad 0 \quad I_{n+1}\}$$

and U, Y has the same definitions as those in Theorem 1.

Proof. Let

$$N = \begin{bmatrix} U^t & -Y^t \\ -Y & U \end{bmatrix}, \quad N_r = \begin{bmatrix} U^t & -Y_r^t \\ -Y_r & U \end{bmatrix} \quad (37)$$

It is obvious that N, N_r are invertible if and only if $U^t U - Y^t Y, U^t U - Y_r^t Y_r$ are invertible.

Since $\max_{|t| < \kappa} \sigma(Y_r, U^{-1}) < 1$ by assumption, we claim that $U^t U - Y_r^t Y_r$ is invertible and $U^t U - Y^t Y$ is invertible for all Y .

On the other hand, from the property of matrices U, Y, Y_r , that is, $Y U = Y_r U$ are commutative [38], it can be proved that

$$N^{-1} = \begin{bmatrix} U(U^t U - Y^t Y)^{-1} & Y^t(U U^t - Y Y^t)^{-1} \\ Y(U^t U - Y^t Y)^{-1} & U(U U^t - Y Y^t)^{-1} \end{bmatrix}, \quad (38)$$

$$N_r^{-1} = \begin{bmatrix} U(U^t U - Y_r^t Y_r)^{-1} & Y_r^t(U U^t - Y_r Y_r^t)^{-1} \\ Y_r(U^t U - Y_r^t Y_r)^{-1} & U(U U^t - Y_r Y_r^t)^{-1} \end{bmatrix} \quad (39)$$

Define

$$M = \begin{bmatrix} U \\ Y \end{bmatrix} (U^t U - Y^t Y)^{-1}, \quad M_r = \begin{bmatrix} U \\ Y_r \end{bmatrix} (U^t U - Y_r^t Y_r)^{-1}. \quad (40)$$

Then, it is clear

$$M = N^{-1} \begin{bmatrix} I_{n+1} \\ 0 \end{bmatrix}, \quad M_r = N_r^{-1} \begin{bmatrix} I_{n+1} \\ 0 \end{bmatrix}. \quad (41)$$

On the other hand, we have

$$\begin{aligned} N_r^{-1} &= \left[\begin{bmatrix} U^t & -Y^t \\ -Y & U \end{bmatrix} + \begin{bmatrix} 0 & (Y - Y_r)^t \\ Y - Y_r & 0 \end{bmatrix} \right]^{-1} \\ &= \left[I_{2(n+1)} + N^{-1} \begin{bmatrix} 0 & E^t \\ E & 0 \end{bmatrix} \right]^{-1} N^{-1} \\ &= N^{-1} - N^{-1} \begin{bmatrix} 0 & E^t \\ E & 0 \end{bmatrix} [I_{2(n+1)}] \\ &\quad + N^{-1} \begin{bmatrix} 0 & E^t \\ E & 0 \end{bmatrix}^{-1} N^{-1} \end{aligned} \quad (42)$$

in which

$$E = Y - Y_r = \begin{bmatrix} v_0 & 0 & \cdots & 0 \\ v_1 & v_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ v_n & v_{n-1} & \cdots & v_0 \end{bmatrix}. \quad (43)$$

Therefore

$$\begin{aligned} M_r &= N_r^{-1} \begin{bmatrix} I_{n+1} \\ 0 \end{bmatrix} \\ &= \left\{ N^{-1} - N^{-1} \begin{bmatrix} 0 & E^t \\ E & 0 \end{bmatrix} \right. \\ &\quad \cdot \left. \left[I_{2(n+1)} + N^{-1} \begin{bmatrix} 0 & E^t \\ E & 0 \end{bmatrix} \right]^{-1} N^{-1} \right\} \begin{bmatrix} I_{n+1} \\ 0 \end{bmatrix} \\ &= M - N^{-1} \begin{bmatrix} 0 & E^t \\ E & 0 \end{bmatrix} [I_{2(n+1)}] \\ &\quad + N^{-1} \begin{bmatrix} 0 & E^t \\ E & 0 \end{bmatrix}^{-1} M. \end{aligned} \quad (44)$$

Let $\delta_i = \frac{1}{\kappa} v_i$, $0 \leq i \leq n$. Since $v_i \in \mathcal{R}$, $|v_i| \leq \kappa$, we have $\delta_i \in \mathcal{R}$, $|\delta_i| \leq 1$, $0 \leq i \leq n$.

From the structure of matrix E , we have

$$\begin{aligned} E^t &= \begin{bmatrix} v_0 & v_1 & \cdots & v_n \\ 0 & v_0 & \cdots & v_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_0 \end{bmatrix} = \sum_{i=0}^n v_i T_2^i \\ &= [I_{n+1} \quad T_2 \quad T_2^2 \quad \cdots \quad T_2^n] \\ &\quad \cdot \begin{bmatrix} \frac{v_0}{\kappa} I_{n+1} & & & \\ & \frac{v_1}{\kappa} I_{n+1} & & \\ & & \ddots & \\ & & & \frac{v_n}{\kappa} I_{n+1} \end{bmatrix} \begin{bmatrix} \kappa I_{n+1} \\ \kappa I_{n+1} \\ \vdots \\ \kappa I_{n+1} \end{bmatrix} \\ &= E_l \tilde{\Lambda} E_r. \end{aligned} \quad (45)$$

Here, T_2^0 is defined as I_{n+1} , and $\tilde{\Lambda} = \text{diag}\{\delta_0 I_{n+1} \quad \delta_1 I_{n+1} \quad \cdots \quad \delta_n I_{n+1}\}$.

Therefore

$$\begin{aligned} \begin{bmatrix} 0 & E^t \\ E & 0 \end{bmatrix} &= \begin{bmatrix} 0 & E_l \tilde{\Lambda} E_r \\ E_r^t \tilde{\Lambda} E_l^t & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & E_l \\ E_r^t & 0 \end{bmatrix} \begin{bmatrix} \tilde{\Lambda} & 0 \\ 0 & \tilde{\Lambda} \end{bmatrix} \begin{bmatrix} E_l^t & 0 \\ 0 & E_r \end{bmatrix}. \end{aligned} \quad (46)$$

Note that

$$\begin{aligned} \begin{bmatrix} \tilde{\Lambda} & 0 \\ 0 & \tilde{\Lambda} \end{bmatrix} &= \text{diag}\{\delta_0 I_{n+1} \quad \delta_1 I_{n+1} \quad \cdots \\ &\quad \delta_n I_{n+1} \quad \delta_0 I_{n+1} \quad \delta_1 I_{n+1} \quad \cdots \quad \delta_n I_{n+1}\} \\ &= \delta_0 \text{diag}\{I_{n+1} \quad 0 \quad \cdots \quad 0 \quad I_{n+1} \quad 0 \quad \cdots \quad 0\} \\ &\quad + \delta_1 \text{diag}\{0 \quad I_{n+1} \quad 0 \quad \cdots \quad 0 \quad I_{n+1} \quad 0 \quad \cdots \quad 0\} \\ &\quad + \cdots + \delta_n \text{diag}\{0 \quad \cdots \quad 0 \quad I_{n+1} \quad 0 \quad \cdots \quad I_{n+1}\} \\ &= [I_{2(n+1)}] \cdot \text{diag}\{\delta_0 I_{2(n+1)} \quad \delta_1 I_{2(n+1)} \quad \cdots \\ &\quad \delta_n I_{2(n+1)}\} [\tilde{S}_0 \quad \tilde{S}_1 \quad \cdots \quad \tilde{S}_n]^t \\ &= \tilde{T}_l \tilde{\Lambda} \tilde{T}_r, \end{aligned} \quad (47)$$

we have

$$\begin{aligned} \begin{bmatrix} 0 & E^t \\ E & 0 \end{bmatrix} &= \begin{bmatrix} 0 & E_l \\ E_r^t & 0 \end{bmatrix} \tilde{T}_l \tilde{\Lambda} \tilde{T}_r \begin{bmatrix} E_l^t & 0 \\ 0 & E_r \end{bmatrix} \\ &= W_l \tilde{\Lambda} W_r, \end{aligned} \quad (48)$$

Therefore

$$\begin{aligned} M_r &= M - N^{-1} W_l \tilde{\Lambda} W_r [I_{2(n+1)} + N^{-1} W_l \tilde{\Lambda} W_r]^{-1} M \\ &= M - N^{-1} W_l \tilde{\Lambda} [I_{2(n+1)} + W_r N^{-1} W_l \tilde{\Lambda}]^{-1} W_r M \end{aligned} \quad (49)$$

Since

$$\begin{bmatrix} p \\ \hat{q} \end{bmatrix} = \begin{bmatrix} U(U^t U - Y^t Y)^{-1} \\ Y(U^t U - Y^t Y)^{-1} \end{bmatrix} = M \epsilon, \quad (50)$$

$$\begin{bmatrix} \hat{p} \\ \hat{q} \end{bmatrix} = \begin{bmatrix} U(U^t U - Y_r^t Y_r)^{-1} \\ Y_r(U^t U - Y_r^t Y_r)^{-1} \end{bmatrix} = M_r \epsilon \quad (51)$$

we conclude

$$\begin{aligned} \begin{bmatrix} \hat{p} \\ \hat{q} \end{bmatrix} &= \begin{bmatrix} p \\ \hat{q} \end{bmatrix} - \begin{bmatrix} U^t & -Y^t \\ -Y & U \end{bmatrix}^{-1} W_l \tilde{\Lambda} \\ &\quad \cdot \left[I_{2(n+1)} + W_r \begin{bmatrix} U^t & -Y^t \\ -Y & U \end{bmatrix}^{-1} W_l \tilde{\Lambda} \right]^{-1} W_r \begin{bmatrix} \hat{p} \\ \hat{q} \end{bmatrix}. \end{aligned} \quad (52)$$

This completes the proof. ■

Theorem 2 suggests that when there exists magnitude bounded noise in the measured plant output, then, if the noise is not too significant, the parameters that describe the possible plant transfer function can be represented by a linear fractional transformation of two matrices. One of them is completely determined by the measured plant input-output pairs and the noise magnitude and is therefore fixed, another one is free, except that its structure is fixed and its maximum singular value is bounded by one.

Let

$$\Phi(z) = \begin{bmatrix} \sum_{j=1}^{n+1} \hat{p}_j^* z^{n+2-j} & \sum_{j=1}^{n+1} \hat{q}_j^* z^{j-1} \\ \sum_{j=1}^{n+1} \hat{q}_j^* z^{n+2-j} & \sum_{j=1}^{n+1} \hat{p}_j^* z^{j-1} \end{bmatrix}$$

then, according to Theorem 1 and the discussion followed it, a transfer function, say, $q(z)$, which has \mathcal{H}^∞ -norm not greater than 1 and matches the input-output data $(u_i, y_i)_{i=0}^n$, has the form

$$q(z) = \mathcal{HM}(\Phi(z), c(z)), \quad c(z) \in \mathcal{BH}^\infty$$

provided that $Y_t^t Y_t \leq U^t U$

Let

$$Z_1 = [z^{n+1} \ z^n \ \dots \ z]^t, \quad Z_2 = [1 \ z \ \dots \ z^n]^t$$

we have

$$\Phi(z) = \begin{bmatrix} \hat{p}^{rt} Z_1 & \hat{q}^{rt} Z_2 \\ \hat{q}^{rt} Z_1 & \hat{p}^{rt} Z_2 \end{bmatrix} = \begin{bmatrix} \hat{p}^{rt} & q^{rt} \\ q^{rt} & \hat{p}^{rt} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \quad (53)$$

On the other hand

$$\begin{aligned} [p^{rt} \ q^{rt}] &= \begin{bmatrix} p^t \\ q^t \end{bmatrix}^t \\ &= \left\{ \begin{bmatrix} p \\ q \end{bmatrix} - \begin{bmatrix} U^t & -Y^t \\ -Y & U \end{bmatrix}^{-1} W_t \bar{\Lambda} \right. \\ &\quad \left. \begin{bmatrix} I_{2(n+1)} \\ I_{2(n+1)} \end{bmatrix} \right. \\ &\quad \left. + W_t \begin{bmatrix} U^t & -Y^t \\ -Y & U \end{bmatrix}^{-1} W_t \bar{\Lambda} \right\}^t \\ &= [p^t \ q^t] - [p^t \ q^t] W_t^t \begin{bmatrix} I_{2(n+1)} \\ I_{2(n+1)} \end{bmatrix} + \bar{\Lambda} W_t^t \\ &\quad \begin{bmatrix} U^t & -Y^t \\ -Y & U \end{bmatrix}^{-1} W_t^t \\ &\quad \bar{\Lambda} W_t^t \begin{bmatrix} U & -Y \\ -Y & U \end{bmatrix}^{-1} \\ &= [p^t \ q^t] - [p^t \ q^t] W_t^t \bar{\Lambda} [I_{2(n+1)} + W_t^t \\ &\quad \begin{bmatrix} U & -Y \\ -Y & U \end{bmatrix}^{-1} W_t^t \bar{\Lambda}]^{-1} \\ &\quad W_t^t \begin{bmatrix} U & Y \\ -Y & U \end{bmatrix}^{-1} \end{aligned} \quad (54)$$

Let

$$I_3 = \begin{bmatrix} 0 & I_{n+1} \\ I_{n+1} & 0 \end{bmatrix}$$

then

$$[q^{rt} \ \hat{p}^{rt}] = [\hat{p}^{rt} \ q^{rt}] T_1 \quad (55)$$

Therefore

$$\begin{aligned} \begin{bmatrix} p^{rt} & q^{rt} \\ q^{rt} & \hat{p}^{rt} \end{bmatrix} &= \begin{bmatrix} N_{11} & 0 \\ 0 & N_{11} \end{bmatrix} \begin{bmatrix} I_{2(n+1)} \\ I_3 \end{bmatrix} \\ &\quad + \begin{bmatrix} N_{12} & 0 \\ 0 & N_{12} \end{bmatrix} \begin{bmatrix} \bar{\Lambda} & 0 \\ 0 & \bar{\Lambda} \end{bmatrix} \begin{bmatrix} I_{4(n+1)} \\ I_3 \end{bmatrix} \\ &\quad - \begin{bmatrix} N_{22} & 0 \\ 0 & N_{22} \end{bmatrix} \begin{bmatrix} \bar{\Lambda} & 0 \\ 0 & \bar{\Lambda} \end{bmatrix}^{-1} \\ &\quad \begin{bmatrix} N_{21} & 0 \\ 0 & N_{21} \end{bmatrix} \begin{bmatrix} I_{2(n+1)} \\ I_3 \end{bmatrix} \end{aligned} \quad (56)$$

Here, $N_{11} = [\hat{p}^t \hat{q}^t]$, $N_{12} = -[\hat{p}^t \hat{q}^t] W_t^t$, $N_{21} = W_t^t$, $\begin{bmatrix} -U & Y^t \\ Y & -U^t \end{bmatrix}^{-1}$, and $N_{22} = W_t^t \begin{bmatrix} U & -Y^t \\ -Y & U^t \end{bmatrix}^{-1} W_t^t$

On the other hand, a simple computation shows that

$$\begin{aligned} \begin{bmatrix} \bar{\Lambda} & 0 \\ 0 & \bar{\Lambda} \end{bmatrix} &= diag\{\delta_0 I_{2(n+1)}, \delta_1 I_{2(n+1)}, \\ &\delta_n I_{2(n+1)}, \delta_0 I_{2(n+1)}, \delta_1 I_{2(n+1)}, \delta_n I_{2(n+1)}\} \\ &= T_1 \Lambda T_1 \end{aligned}$$

in which

$$\Lambda = diag\{\delta_0 I_{4(n+1)}, \delta_1 I_{4(n+1)}, \delta_n I_{4(n+1)}\}, \quad (57)$$

$$T_1 = [I_{4(n+1)}, I_{4(n+1)}, I_{4(n+1)}], \quad (58)$$

$$I_1 = [S_0 \ S_1 \ \dots \ S_n]^t \quad (59)$$

$$S_0 = diag\{I_{2(n+1)} \ 0 \ 0 \ I_{2(n+1)} \ 0 \ 0\}, \quad (60)$$

$$S_1 = diag\{0 \ I_{2(n+1)} \ 0 \ 0 \ I_{2(n+1)} \ 0 \ 0\} \quad (61)$$

$$S_n = diag\{0 \ 0 \ I_{2(n+1)} \ 0 \ 0 \ I_{2(n+1)}\} \quad (62)$$

Let

$$R_{11} = \begin{bmatrix} N_{11} & 0 \\ 0 & N_{11} \end{bmatrix} \begin{bmatrix} I_{2(n+1)} \\ I_1 \end{bmatrix}, \quad R_{12} = \begin{bmatrix} N_{12} & 0 \\ 0 & N_{12} \end{bmatrix},$$

$$R_{21} = \begin{bmatrix} N_{21} & 0 \\ 0 & N_{21} \end{bmatrix} \begin{bmatrix} I_{2(n+1)} \\ I_1 \end{bmatrix}, \quad R_{22} = \begin{bmatrix} N_{22} & 0 \\ 0 & N_{22} \end{bmatrix}$$

Then

$$\begin{aligned} \begin{bmatrix} p^{rt} & q^{rt} \\ q^{rt} & \hat{p}^{rt} \end{bmatrix} &= R_{11} + R_{12} T_1 \Lambda T_1 [I_{4(n+1)} - R_{22} T_1 \Lambda]^{-1} R_{21} \\ &= R_{11} + R_{12} I_1 \Lambda [I_{4(n+1)} - T_1 R_{22} T_1 \Lambda]^{-1} I_1 R_{21} \end{aligned} \quad (63)$$

Therefore

$$\begin{aligned} \Phi(z) &= \begin{bmatrix} p^{rt} & q^{rt} \\ q^{rt} & \hat{p}^{rt} \end{bmatrix} \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \\ &= \{R_{11} + R_{12} T_1 \Lambda [I_{4(n+1)} - T_1 R_{22} T_1 \Lambda]^{-1} T_1 R_{21}\} \\ &\quad \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \\ &= W_{11}(z) + W_{12} \Lambda [I_{4(n+1)} - W_{22} \Lambda]^{-1} W_{21}(z) \end{aligned} \quad (64)$$

Here

$$W_{11}(z) = R_{11} \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}, \quad W_{12} = R_{12} T_1,$$

$$W_{21}(z) = T_1 R_{21} \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}, \quad W_{22} = T_1 R_{22} T_1$$

From the construction, it is obvious that $W_{11}(z)$ is a transfer function matrix belonging to \mathcal{H}^∞ with dimension 2×2 , W_{12}

is a matrix belonging to $\mathcal{R}^{2 \times 4(n+1)^4}$, $W_{21}(z)$ is a transfer function matrix belonging to \mathcal{H}^∞ with dimension $4(n+1)^4 \times 2$, while matrix W_{22} belongs to $\mathcal{R}^{4(n+1)^4 \times 4(n+1)^4}$.

From the above discussion, the following theorem is obtained.

Theorem 3: Let

$$W(z) = \begin{bmatrix} W_{11}(z) & W_{12} \\ W_{21}(z) & W_{22} \end{bmatrix}$$

then, $\Phi(z) = \mathcal{F}_l(W(z), \Lambda)$.

The above theorem makes it clear that when there exists magnitude bounded noise in the measured plant output, and the noise is not too significant, the homographic transformation that describes all of the plant possible transfer functions, can be parameterized as a linear fractional transformation of a fixed transfer function matrix and a structure fixed, norm bounded, but free matrix.

The next theorem is the main results of this paper.

Theorem 4: Partition $W_{11}(z)$, W_{12} , $W_{21}(z)$ as

$$W_{11}(z) = \begin{bmatrix} W_{11}^{11}(z) & W_{11}^{12}(z) \\ W_{11}^{21}(z) & W_{11}^{22}(z) \end{bmatrix}, \quad W_{12} = \begin{bmatrix} W_{12}^{11} \\ W_{12}^{21} \end{bmatrix},$$

$$W_{21}(z) = [W_{21}^{11}(z) \quad W_{21}^{12}(z)]$$

in which $W_{11}^{11}(z)$ has a dimension 1×1 , W_{12}^{11} has a dimension $1 \times 4(n+1)^4$, $W_{21}^{11}(z)$ has a dimension $4(n+1)^4 \times 1$, and the others have compatible dimensions. Moreover, see the matrices at the bottom of the page. If $\max_{|z| \leq \kappa} \bar{\sigma}((Y + E)U^{-1}) < 1$, then, transfer function set

$$\mathcal{G} = \left\{ g(z) \mid g(z) = \mathcal{F}_l(F(z), \Delta(z)), \max_{|z| \leq 1} \bar{\sigma}(\Delta(z)) \leq 1 \right\}$$

exhausts all of the transfer functions that have \mathcal{H}^∞ -norm not greater than 1 and match the possible plant input-output pairs that are derived from the measured plant input-output data $(u_i, y_i) \mid_{i=0}^n$ which are contaminated by a magnitude bounded noise series v_i , $v_i \in \mathcal{R}$, $|v_i| \leq \kappa$, $0 \leq i \leq n$ in y_i in additive form, and vice versa. Here, the matrices U , Y are the same as those defined in Theorem 2, and the matrices E , Λ are defined by (43) and (57), respectively.

Proof: From Theorem 2 and Theorem 3, it is obvious that transfer function set

$$\bar{\mathcal{G}} = \{g(z) \mid g(z) = \mathcal{H}\mathcal{M}(\mathcal{F}_l(W(z), \Lambda), \epsilon(z)),$$

$$b_i \in \mathcal{R}, \bar{\sigma}(\Lambda) \leq 1, \epsilon(z) \in \mathcal{B}\mathcal{H}^\infty\} \quad (65)$$

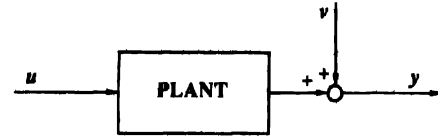


Fig. 1. Identification experiment.

equals the transfer function set which includes all of the transfer functions that do not contradict the plant available information.

Note that every transfer function in $\bar{\mathcal{G}}$ has a structure like that depicted in Fig. 2.

The internal relations of Fig. 2 are

$$\begin{bmatrix} y(z) \\ u(z) \\ w_2(z) \end{bmatrix} = \begin{bmatrix} q_{11}(z) & q_{12}(z) & q_{13}(z) \\ q_{21}(z) & q_{22}(z) & q_{23}(z) \\ q_{31}(z) & q_{32}(z) & q_{33}(z) \end{bmatrix} \begin{bmatrix} v_1(z) \\ w_1(z) \\ v_2(z) \end{bmatrix}, \quad (66)$$

$$v_1(z) = \epsilon(z)w_1(z), \quad (67)$$

$$v_2(z) = \Lambda w_2(z). \quad (68)$$

Note that

$$q_{22}(z) = W_{11}^{22}(z) = \bar{p}' Z_2 \quad (69)$$

which is invertible on the closed unit disk $D = \{z \mid |z| \leq 1\}$.

On the other hand, the relation in (66) can be rewritten as

$$\begin{bmatrix} 1 & -q_{12}(z) & 0 \\ 0 & -q_{22}(z) & 0 \\ 0 & -q_{32}(z) & I \end{bmatrix} \begin{bmatrix} y(z) \\ w_1(z) \\ w_2(z) \end{bmatrix} = \begin{bmatrix} 0 & q_{11}(z) & q_{13}(z) \\ -1 & q_{21}(z) & q_{23}(z) \\ 0 & q_{31}(z) & q_{33}(z) \end{bmatrix} \begin{bmatrix} u(z) \\ v_1(z) \\ v_2(z) \end{bmatrix}. \quad (70)$$

Since

$$\begin{bmatrix} 1 & -q_{12}(z) & 0 \\ 0 & -q_{22}(z) & 0 \\ 0 & -q_{32}(z) & I \end{bmatrix}$$

is invertible if and only if $q_{22}(z)$ is invertible. (70) equals (71) (found at the bottom of the next page) which implies that

$$\mathcal{G} = \bar{\mathcal{G}}. \quad (72)$$

$$Q(z) = \begin{bmatrix} q_{11}(z) & q_{12}(z) & q_{13}(z) \\ q_{21}(z) & q_{22}(z) & q_{23}(z) \\ q_{31}(z) & q_{32}(z) & q_{33}(z) \end{bmatrix} = \begin{bmatrix} W_{11}^{11}(z) & W_{11}^{12}(z) & W_{11}^{13} \\ W_{11}^{21}(z) & W_{11}^{22}(z) & W_{11}^{23} \\ W_{21}^{11}(z) & W_{21}^{12}(z) & W_{22} \end{bmatrix},$$

$$F(z) = \begin{bmatrix} q_{12}(z)q_{22}^{-1}(z) & q_{11}(z) - q_{12}(z)q_{22}^{-1}(z)q_{21}(z) & q_{13}(z) - q_{12}(z)q_{22}^{-1}(z)q_{23}(z) \\ q_{22}^{-1}(z) & -q_{22}^{-1}(z)q_{21}(z) & -q_{22}^{-1}(z)q_{23}(z) \\ q_{32}(z)q_{22}^{-1}(z) & q_{31}(z) - q_{32}(z)q_{22}^{-1}(z)q_{21}(z) & q_{33}(z) - q_{32}(z)q_{22}^{-1}(z)q_{23}(z) \end{bmatrix},$$

$$\Delta(z) = \begin{bmatrix} \epsilon(z) & 0 \\ 0 & \Lambda \end{bmatrix}.$$

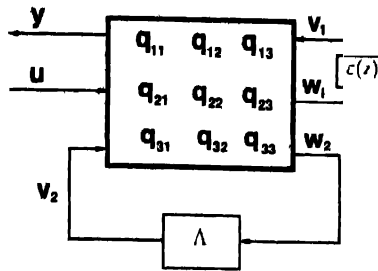
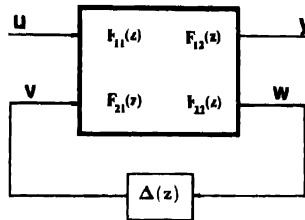
Fig. 2 Structure of transfer function belonging to $\bar{\mathcal{G}}$ 

Fig. 3 Structure of plant possible transfer function

This completes the proof. ■

From Theorem 4 we see that the transfer function set which includes all of the plant possible transfer functions can be parameterized by a linear fractional transformation of a fixed transfer function matrix and a structure fixed norm bounded, free transfer function matrix.

The structure of the possible plant transfer function is illustrated in Fig. 3.

The results of Theorem 4 can be extended to the case when $\max_{|z| \leq \kappa} \bar{\sigma}[(I + E)U^{-1}] \leq 1$. In this situation all the plant possible transfer functions can also be parameterized as \mathcal{G} in Theorem 4. The idea behind the proof of this conclusion is to firstly parameterize all the possible transfer functions for the plant whose transfer function's \mathcal{H}^∞ norm is not greater than $1 + \alpha$ for arbitrary $\alpha > 0$ and whose input-output pairs can be deduced from the measured input-output data $(u_i, y_i) |_{i=0}^n$, represent it as $\mathcal{G}(1 + \alpha)$ and then, take the limit of the transfer function set $\mathcal{G}(1 + \alpha)$ as α tends to 0 from the positive side. Based on the relations that $\mathcal{G}(1 + \alpha) \supset \mathcal{G}$ for any $\alpha > 0$, and $\lim_{\alpha \rightarrow 0^+} \mathcal{G}(1 + \alpha) = \mathcal{G}$ we obtain the transfer function set with the desirable properties. The proof is elementary but tedious, and is therefore omitted here. The details of the proof can be found in [41].

When the noise magnitude bound κ does not satisfy the condition mentioned above, i.e., $\max_{|z| \leq \kappa} \bar{\sigma}[(I + E)U^{-1}] \leq 1$, let

$$\nabla = \begin{bmatrix} \delta_0 & 0 & 0 \\ \delta_1 & \delta_0 & 0 \\ \vdots & \vdots & \vdots \\ \delta_n & \delta_{n-1} & \delta_0 \end{bmatrix}$$

For those $\delta_0, \delta_1, \dots, \delta_n$, such that $(\frac{1}{\kappa}Y + \nabla)^t (\frac{1}{\kappa}Y + \nabla) > \frac{1}{\kappa}U^t U$, or $(Y + \kappa \nabla)^t (Y + \kappa \nabla) > U^t U$, and $\delta_i \in \mathcal{R}$, $|\delta_i| \leq 1$, the plant under investigation cannot have a response $(y_i + \kappa \delta_i) |_{i=0}^n$ when the input series $u_i |_{i=0}^n$ is applied to it, since according to the input-output extrapolation-minimization theorem [28], [36]–[37], every transfer function that can match these input-output pairs, i.e., $(u_i, y_i + \kappa \delta_i) |_{i=0}^n$, will have a \mathcal{H}^∞ -norm greater than 1. Videlicet, these noise series $(-\kappa \delta_i) |_{i=0}^n$ are impossible according to the plant assumptions. Therefore, they should be excluded from the noise set $\mathcal{V} = \{v_i |_{i=0}^n | v_i \in \mathcal{R}, |v_i| \leq \kappa\}$. From the above discussion, it is apparent that in this case the transfer function set that contains all of the plant possible transfer functions equals transfer function set

$$\mathcal{G} = \{g(\cdot) | g(\cdot) = \mathcal{F}_l(F(\cdot), \Delta(\cdot))\}$$

$$\max_{|\delta_i| \leq 1} \bar{\sigma}(\Delta(\cdot)) \leq 1 \quad (\delta_0, \delta_1, \dots, \delta_n) \in D_\delta$$

Here $D_\delta = \{(\delta_0, \delta_1, \dots, \delta_n) | (\frac{1}{\kappa}Y + \nabla)^t (\frac{1}{\kappa}Y + \nabla) \leq \frac{1}{\kappa}U^t U, \delta_i \in \mathcal{R}, |\delta_i| \leq 1\}$, and the other matrices or transfer function matrices have the same definitions as those in Theorem 4.

IV APPLICATION TO IDENTIFICATION FOR ROBUST CONTROL

In Section III we showed that when the magnitude bounded additive noise existent in the measure plant output is not too significant the transfer function set that consists of all plant possible transfer functions can be parameterized as a linear fractional transformation of a fixed transfer function matrix and a structure fixed norm bounded, free transfer function matrix. The importance of linear fractional transformation has been widely recognized in control community in recent years, as well as homographic transformation [5], [11], [20], [27]. Note that all of the transfer function set descriptions widely used in robust control theory, such as nominal model

$$\begin{aligned} \begin{bmatrix} y(z) \\ w_1(z) \\ w_2(z) \end{bmatrix} &= \begin{bmatrix} 1 & -q_{12}(z) & 0 \\ 0 & -q_{22}(z) & 0 \\ 0 & -q_{32}(z) & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & q_{11}(z) & q_{13}(z) \\ -1 & q_{21}(z) & q_{23}(z) \\ 0 & q_{31}(z) & q_{33}(z) \end{bmatrix} \begin{bmatrix} u(z) \\ v_1(z) \\ v_2(z) \end{bmatrix} \\ &= \begin{bmatrix} q_{12}(z)q_{22}^{-1}(z) & q_{11}(z) - q_{12}(z)q_{22}^{-1}(z)q_{21}(z) & q_{13}(z) - q_{12}(z)q_{22}^{-1}(z)q_{23}(z) \\ q_{22}^{-1}(z) & -q_{22}^{-1}(z)q_{21}(z) & -q_{22}^{-1}(z)q_{23}(z) \\ q_{32}(z)q_{22}^{-1}(z) & q_{31}(z) - q_{32}(z)q_{22}^{-1}(z)q_{21}(z) & q_{33}(z) - q_{32}(z)q_{22}^{-1}(z)q_{23}(z) \end{bmatrix} \begin{bmatrix} u(z) \\ v_1(z) \\ v_2(z) \end{bmatrix} \\ &= F(z) \begin{bmatrix} u(z) \\ v_1(z) \\ v_2(z) \end{bmatrix} \end{aligned} \quad (71)$$

with parametric uncertainty bound, nominal model with additive unstructured uncertainty bound, nominal model with multiplicative unstructured uncertainty bound, nominal model with coprime factor unstructured uncertainty bound, etc., can be easily transformed into a linear fractional transformation representation, which conversely implies that linear fractional transformation with structure fixed, norm bounded, uncertain transfer function matrix is in fact one of the most general descriptions of plant model with intended application as robust controller design. Moreover, a transfer function set, which is parameterized by linear fractional transformation of a fixed transfer function matrix and a norm bounded, structure fixed, uncertain transfer function matrix, can be directly used in robust controller design, because such a parameterization just fits the framework of structured singular value theory or μ , although this robust controller design theory is still not complete and more powerful design algorithms are needed to be developed [4], [8], [27].

In controller design, however, the direct utilization of the transfer function set given in Section III suffers from several drawbacks. For instance, the plant nominal model generally has a high degree and the number of the uncertain parameters is large. In fact, they generally equal to the identification experiment data length and the identification experiment data length + 1, respectively. Here, by nominal model, we mean the transfer function corresponding to $\Delta(z) = 0$. Moreover, the dimension of the matrices used in the parameterization is extremely large, e.g., the dimension of the uncertain transfer function matrix is $4(n+1)^4 + 1$, in which n stands for the identification experiment data length. But in controller design, it is preferable to use low complexity nominal model and a simple description of the nominal model error uncertainty bound, because such a plant model will represent the dominant characteristics of the plant dynamics which is more convenient in control system analysis and synthesis, and will lead to a low complexity controller which is highly appreciated in realization. Furthermore, we note that in the plant possible transfer function parameterization, there are both dynamic uncertainty and real perturbation. The dynamic uncertainty is due to the plant unmeasured output, while the real perturbation is due to the effect of noise. If this parameterization is directly applied to robust controller design, it will lead to a mixed μ -synthesis problem, which is well known as NP hard [8], [27].

In this section, we would like to briefly discuss the problem of obtaining the desirable fixed complexity plant nominal model and its error uncertainty bound from the aforementioned plant available information.

First, we introduce some notations that will be used in this section.

For a given matrix structure \mathcal{X} , let \mathcal{BX} represent the unit ball defined as $\mathcal{BX} = \{\Delta \mid \Delta \in \mathcal{X}, \bar{\sigma}(\Delta) \leq 1\}$. When $\mathcal{X}(z)$ is a transfer function matrix structure, we define $\mathcal{BX}(z)$ as $\mathcal{BX}(z) = \{\Delta(z) \mid \Delta(z) \in \mathcal{X}(z), \max_{|z| \leq 1} \bar{\sigma}(\Delta(z)) \leq 1\}$. Moreover, for a given matrix structure \mathcal{X} and a transfer function matrix $H(z)$ which belongs to \mathcal{H}^∞ and has suitable dimension, with a little abuse of notation, we define $\mu_{\mathcal{X}}(H)$ as the maximal structured singular value of $H(z)$ according

to \mathcal{X} over the closed unit disk $\mathcal{D} = \{z \mid |z| \leq 1\}$, i.e., $\mu_{\mathcal{X}}(H) := \max_{|z| \leq 1} \mu_{\mathcal{X}}(H(z))$.

Note that for a transfer function matrix $H(z') = D + C(z'I - A)^{-1}B$ which is analytic on the complement of the open unit disk, i.e., $\mathcal{D}_{open}^c = \{z' \mid |z'| \geq 1\}$ ² let $z = 1/z'$, then, $H(z')$ can be rewritten as $H(z) = \mathcal{F}_l\left(\begin{bmatrix} D & C \\ B & A \end{bmatrix}, zI\right)$. It is obvious that $H(z) \in \mathcal{H}^\infty$ and from the main loop theorem that plays an important role in μ -analysis and μ -synthesis [27], it can be proved that $\mu_{\mathcal{X}}(H) \leq \beta$ for a positive real number β if and only if $\mu_{\overline{\mathcal{X}}}(M) \leq 1$. Here, $M = \begin{bmatrix} D/\beta & C/\sqrt{\beta} \\ B/\sqrt{\beta} & A \end{bmatrix}$, and $\overline{\mathcal{X}}$ is a matrix structure defined as $\overline{\mathcal{X}} = \{\overline{\Delta} \mid \overline{\Delta} = \text{diag}\{\Delta, \epsilon I\}, \Delta \in \mathcal{X}, \epsilon \in \mathcal{C}\}$. Especially, when $\mu_{\mathcal{X}}(H) = 1$, it can be shown that $\mu_{\overline{\mathcal{X}}}(M) = 1$, also. To express the results in this section more concisely, we adopt the notation $\mu_{\mathcal{X}}(H)$ for a transfer function matrix $H(z) \in \mathcal{H}^\infty$ and a compatible matrix structure \mathcal{X} .

Using a similar argument as that in the proof of the main loop theorem, we can obtain the following lemma.

Lemma 1: For a $(m+p) \times (m+p)$ transfer function matrix $H(z) = \begin{bmatrix} H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z) \end{bmatrix} \in \mathcal{H}^\infty$, a matrix structure \mathcal{X}_1 with dimension $m \times m$, and a matrix structure \mathcal{X}_2 with dimension $p \times p$, define a new matrix structure \mathcal{X}' as $\mathcal{X}' = \left\{ \Delta = \begin{bmatrix} \Delta_1 & \Delta_2 \end{bmatrix} \mid \Delta_1 \in \mathcal{X}_1, \Delta_2 \in \mathcal{X}_2 \right\}$. If $\mu_{\mathcal{X}_2}(H_{22}) < 1$, then, $\mu_{\mathcal{X}'}(H) \leq 1$, if and only if $\max_{\Delta_2 \in \mathcal{BX}_2} \mu_{\mathcal{X}_1}(\mathcal{F}_l(H, \Delta_2)) \leq 1$.

Proof: Since $\mu_{\mathcal{X}_2}(H_{22}) < 1$, we have eliminate at every $z \in \mathcal{D} = \{z \mid |z| \leq 1\}$

$$\det(I_p - H_{22}(z)\Delta_2) \neq 0 \quad (73)$$

if $\Delta_2 \in \mathcal{BX}_2$. Therefore, for every $z \in \mathcal{D}$, $\Delta_1 \in \mathcal{X}_1$, $\Delta_2 \in \mathcal{BX}_2$, a simple computation will show

$$\begin{aligned} \det\left(I_{m+p} - \begin{bmatrix} H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z) \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix}\right) \\ = \det(I_p - H_{22}(z)\Delta_2) \cdot \det(I_m - \mathcal{F}_l(H(z), \Delta_2)\Delta_1). \end{aligned} \quad (74)$$

Hence, $\mu_{\mathcal{X}'}(H) < 1$ if and only if $\max_{\Delta_2 \in \mathcal{BX}_2} \mu_{\mathcal{X}_1}(\mathcal{F}_l(H, \Delta_2)) < 1$.

If $\max_{\Delta_2 \in \mathcal{BX}_2} \mu_{\mathcal{X}_1}(\mathcal{F}_l(H, \Delta_2)) = 1$, then, there exist at least one $z_0 \in \mathcal{D}$, one $\Delta_2 \in \mathcal{BX}_2$, and one $\Delta_1 \in \mathcal{X}_1$, such that $\bar{\sigma}(\Delta_1) = 1$, and

$$\det(I_m - \mathcal{F}_l(H(z_0), \Delta_2)\Delta_1) = 0 \quad (75)$$

which implies that

$$\det\left(I_{m+p} - \begin{bmatrix} H_{11}(z_0) & H_{12}(z_0) \\ H_{21}(z_0) & H_{22}(z_0) \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix}\right) = 0. \quad (76)$$

Since

$$\bar{\sigma}\left[\begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix}\right] = \max\{\bar{\sigma}(\Delta_1), \bar{\sigma}(\Delta_2)\} = 1 \quad (77)$$

²The analyticity of a transfer function matrix on \mathcal{D}_{open}^c equals the commonly utilized concept of stability of a transfer function matrix in control engineering.

we have

$$\mu_1(H) \leq 1 \quad (78)$$

If $\mu_1(H) = 1$, then, for arbitrary $\alpha > 0$, $\mu_1(H) < 1 + \alpha$. According to the definition of structured singular value, we have that for any $z \in \mathcal{D}$, $\Delta_1 \in \mathcal{X}_1$, $\bar{\sigma}(\Delta_1) \leq \frac{1}{1+\alpha}$, $\Delta_2 \in \mathcal{X}_2$, $\bar{\sigma}(\Delta_2) \leq \frac{1}{1+\alpha}$

$$\det \left(I_{m+p} - \begin{bmatrix} H_{11}(z) & H_{12}(z) \\ H_{21}(z) & H_{22}(z) \end{bmatrix} \begin{bmatrix} \Delta_1 & \\ & \Delta_2 \end{bmatrix} \right) \neq 0 \quad (79)$$

which implies that

$$\max_{\Delta_1 \in \mathcal{X}_1, \Delta_2 \in \mathcal{X}_2} \mu_1(\mathcal{F}_l(H, \Delta_2)) < 1 + \alpha \quad (80)$$

Since α is an arbitrary positive number, we claim that

$$\max_{\Delta_1 \in \mathcal{X}_1, \Delta_2 \in \mathcal{X}_2} \mu_1(\mathcal{F}_l(H, \Delta_2)) \leq 1 \quad (81)$$

This completes the proof. ■

At first we investigate the problem that concerns under what condition a transfer function set described by a nominal model and an additive unstructured uncertainty bound can include all the possible plant transfer functions when an upper bound of the plant transfer function's \mathcal{H}^∞ -norm and a series of the plant's weakly corrupted time domain input-output pairs have been provided.

Partition the transfer function matrix $F(z)$ defined in Theorem 4 as $F(z) = \begin{bmatrix} F_{11}(z) & F_{12}(z) \\ F_{21}(z) & F_{22}(z) \end{bmatrix}$ where $F_{11}(z)$ has a dimension $l \times l$ and the other transfer function matrices have compatible dimensions.

Define transfer function matrix structure $\mathcal{V}(z)$ and matrix structures $\mathcal{X} = \bar{\mathcal{X}}$ as

$$\mathcal{X}(\cdot) = \{ \Delta(\cdot) \mid \Delta(\cdot) = \text{diag} \{ \epsilon(\cdot) \delta_0 I, \delta_1 I, \dots, \delta_n I \} \\ \epsilon(\cdot) \in \mathcal{H}^\infty, \delta_i \in \mathcal{R}, 0 \leq i \leq n \}$$

$$\bar{\mathcal{X}} = \{ \underline{\Delta} \mid \underline{\Delta} = \text{diag} \{ \epsilon \delta_0 I, \delta_1 I, \dots, \delta_n I \} \\ \epsilon \in \mathcal{C}, \delta_i \in \mathcal{R}, 0 \leq i \leq n \}$$

$$\mathcal{V} = \left\{ \bar{\Delta} \mid \bar{\Delta} = \begin{bmatrix} \Delta_1 & \\ & \Delta_2 \end{bmatrix}, \Delta_1 \in \mathcal{C}, \Delta_2 \in \mathcal{V} \right\}$$

Then, according to Theorem 4, when the plant transfer function is derived from an upper bound of its \mathcal{H}^∞ -norm and a series of weakly corrupted time domain input-output data, the transfer function set \mathcal{G} which consists of all the plant possible transfer functions can be parameterized as

$$\mathcal{G} = \{ q(z) \mid q(z) = \mathcal{F}_l(F(z), \Delta(z)), \Delta(z) \in \mathcal{BX}(z) \}$$

Theorem 5 For a given transfer function $q_0(z) \in \mathcal{H}^\infty$ and a unimodular transfer function of \mathcal{H}^∞ , $W_a(z)$, define a transfer function set

$$\mathcal{G}_a = \{ g(z) \mid g(z) = q_0(z) + W_a(z) \Delta_a(z), \Delta_a(z) \in \mathcal{BH}^\infty \} +$$

Then, $\mathcal{G}_a \supseteq \mathcal{G}$ if and only if

$$\mu_{\bar{\mathcal{V}}} \begin{bmatrix} W_a^{-1}(F_{11} - q_0) & W_a^{-1}F_{12} \\ I_{21} & I_{22} \end{bmatrix} \leq 1$$

Proof First, we note that since $\mathcal{G} \subset \mathcal{H}^\infty$, it is necessary that at every $z \in \mathcal{D}$, $(I - F_{22}(z)\Delta(z))$ are invertible, for arbitrary $\Delta(z) \in \mathcal{V}(z)$.

Assume that there exists a $z_0 \in \mathcal{D}$ and a $\underline{\Delta} = \text{diag} \{ \alpha \delta_0 I, \delta_1 I, \dots, \delta_n I \} \in \mathcal{BX}$, such that

$$\det(I - F_{22}(z_0)\underline{\Delta}) = 0 \quad (82)$$

If $|z_0| < 1$ from $|\alpha| \leq 1$ and the Nevanlinna-Pick interpolation theorem [22], we conclude that there exists an $\epsilon(\cdot) \in \mathcal{BH}^\infty$ that satisfies $\epsilon(z_0) = \alpha$. Let $\Delta(z) = \text{diag} \{ \epsilon(z) \delta_0 I, \delta_1 I, \dots, \delta_n I \}$ then, it is obvious that $\Delta(z) \in \mathcal{BX}(\cdot)$ and $\det(I - F_{22}(z_0)\Delta(z_0)) = 0$ which is a contradiction.

If $|z_0| = 1$, let $z'_0 = \beta z_0$ for arbitrary $|\beta| < 1$, then, using a similar argument, it can be proved that, there exists a $\Delta(\cdot) \in \mathcal{BX}(\cdot)$ that meets the condition $\Delta(\beta z_0) = \underline{\Delta}$. From $\Delta(\cdot) \in \mathcal{BX}(\cdot)$, we have

$$\Delta(z_0) = \lim_{\eta \rightarrow 1} \Delta(\eta z_0) \quad (83)$$

Since β is arbitrary except that $|\beta| < 1$, we conclude that there exists a $\Delta(\cdot) \in \mathcal{BX}(\cdot)$ such that

$$\det(I - F_{22}(z_0)\Delta(z_0)) = 0 \quad (84)$$

which is also a contradiction.

Therefore, for every z $|z| \leq 1$, $\det(I - F_{22}(z)\underline{\Delta}) \neq 0$ whenever $\underline{\Delta} \in \mathcal{BX}$ which implies that

$$\mu_{\bar{\mathcal{V}}}(F_{22}) < 1 \quad (85)$$

If $\mathcal{G}_a \supseteq \mathcal{G}$ then for arbitrary transfer function $q(z) \in \mathcal{G}$, there exists a $\Delta_a(z) \in \mathcal{BH}^\infty$ such that

$$q(z) = q_0(z) + W_a(z)\Delta_a(z) \quad (86)$$

That is, there exists at least one $\Delta_a(z) \in \mathcal{BH}^\infty$ which satisfies

$$F_{11}(z) + F_{12}(z)\Delta(z)[I - F_{22}(z)\Delta(z)]^{-1}F_{21}(z) \\ = q_0(z) + W_a(z)\Delta_a(z) \quad (87)$$

for arbitrary $\Delta(z) \in \mathcal{BX}(z)$.

Hence

$$\Delta_a(z) = W_a^{-1}(z)[F_{11}(z) - q_0(z) + F_{12}(z)\Delta(z) \\ (I - F_{22}(z)\Delta(z))^{-1}F_{21}(z)] \\ = \mathcal{F}_l \left[\begin{pmatrix} W_a^{-1}(z)(F_{11}(z) - q_0(z)) & W_a^{-1}(z)F_{12}(z) \\ F_{21}(z) & F_{22}(z) \end{pmatrix} \right. \\ \left. \Delta(z) \right] \quad (88)$$

which implies that

$$\max_{\Delta(z) \in \mathcal{B}\mathcal{X}(z)} \left\| \mathcal{F}_l \left[\begin{pmatrix} W_a^{-1}(z)(F_{11}(z) - g_0(z)) & W_a^{-1}(z)F_{12}(z) \\ F_{21}(z) & F_{22}(z) \end{pmatrix}, \Delta(z) \right] \right\|_{\infty} \leq 1 \quad (89)$$

because $\Delta_a(z) \in \mathcal{B}\mathcal{H}^{\infty}$.

Therefore, using similar arguments as in the proof of $\mu_{\mathcal{X}}(F_{22}) < 1$, we conclude that

$$\max_{\Delta \in \mathcal{B}\mathcal{X}} \mu_c \left(\mathcal{F}_l \left[\begin{pmatrix} W_a^{-1}(F_{11} - g_0) & W_a^{-1}F_{12} \\ F_{21} & F_{22} \end{pmatrix}, \Delta \right] \right) \leq 1. \quad (90)$$

From Lemma 1 and (85) and (90), we have

$$\mu_{\overline{\mathcal{X}}} \left(\begin{pmatrix} W_a^{-1}(F_{11} - g_0) & W_a^{-1}F_{12} \\ F_{21} & F_{22} \end{pmatrix} \right) \leq 1. \quad (91)$$

On the other hand, if

$$\mu_{\overline{\mathcal{X}}} \left(\begin{pmatrix} W_a^{-1}(F_{11} - g_0) & W_a^{-1}F_{12} \\ F_{21} & F_{22} \end{pmatrix} \right) \leq 1$$

then, according to Lemma 1, we have

$$\max_{\Delta \in \mathcal{B}\mathcal{X}} \max_{|z| \leq 1} \left\| \mathcal{F}_l \left[\begin{pmatrix} W_a^{-1}(z)(F_{11}(z) - g_0(z)) & W_a^{-1}(z)F_{12}(z) \\ F_{21}(z) & F_{22}(z) \end{pmatrix}, \Delta \right] \right\|_{\infty} \leq 1. \quad (92)$$

Assume that there is a transfer function $g(z) \in \mathcal{G}$, but $g(z) \notin \mathcal{G}_a$. Since $g(z) \in \mathcal{G}$, there exists a $\Delta(z) \in \mathcal{B}\mathcal{X}(z)$, such that

$$g(z) = F_{11}(z) + F_{12}(z)\Delta(z)(I - F_{22}(z)\Delta(z))^{-1}F_{21}(z). \quad (93)$$

Let $\Delta_a(z) = W_a^{-1}(z)(g(z) - g_0(z))$; then, it is obvious that $\Delta_a(z) \in \mathcal{H}^{\infty}$ and

$$\|\Delta_a(z)\|_{\infty} > 1 \quad (94)$$

which implies that there exists at least one $z_0 \in \mathcal{D}$, at which

$$|\Delta_a(z_0)| > 1. \quad (95)$$

Let $\underline{\Delta} = \Delta(z_0)$, then, it is apparent that

$$\underline{\Delta} \in \mathcal{B}\mathcal{X}. \quad (96)$$

Note that

$$\begin{aligned} \Delta_a(z_0) &= \mathcal{F}_l \left[\begin{pmatrix} W_a^{-1}(z_0)(F_{11}(z_0) - g_0(z_0)) & W_a^{-1}(z_0)F_{12}(z_0) \\ F_{21}(z_0) & F_{22}(z_0) \end{pmatrix}, \underline{\Delta} \right] \\ &\quad \cdot \underline{\Delta} \end{aligned} \quad (97)$$

we conclude that $\|\Delta_a(z)\|_{\infty} > 1$ contradicts with (92).

Therefore, there is no transfer function that belongs to \mathcal{G} but does not belong to \mathcal{G}_a .

This completes the proof. \blacksquare

Noting that unimodular transfer functions of \mathcal{H}^{∞} are capable of representing all the kinds of frequency-magnitude characteristics, the assumption that $W_a(z)$ is a unimodular function of \mathcal{H}^{∞} does not sacrifice any generality concerned with the application of the results in Theorem 5.

From Theorem 5, we see that for a weakly corrupted plant, the problem of checking whether a transfer function set, which is defined by a nominal model and an additive nominal model error uncertainty bound, includes all of the possible plant transfer functions that are deduced from an upper bound of the plant transfer function's \mathcal{H}^{∞} -norm and measured time domain input-output data, is equal to a μ -analysis problem [4], [27]. Here, by weakly corrupted plant, we mean that the normalized measured plant input-output satisfies the condition $\max_{|v_i| \leq \kappa} \overline{\sigma}((Y + E)U^{-1}) \leq 1$.

Finally, we discuss the problem of obtaining the smallest transfer function set which is parameterized by a fixed complexity nominal model with an additive nominal model error uncertainty bound, and includes all the plant possible transfer functions that do not contradict the aforementioned available plant information.

Mathematically, this identification problem can be formulated as finding out the fixed complexity nominal model $g_0(z) \in \mathcal{H}^{\infty}$, such that for a given frequency weight function $W_{ai}(z)$ (which can be chosen as a unimodular transfer function of \mathcal{H}^{∞}) and the unfalsified transfer function set, i.e., \mathcal{G} , the cost function

$$J(g_0(z)) = \max_{g(z) \in \mathcal{G}} \|W_{ai}(z)[g(z) - g_0(z)]\|_{\infty} \quad (98)$$

is minimized. For control engineering background of this problem formulation, refer to [7], [9], [24], and [38].

Let \mathcal{H}_m^{∞} denote the transfer function set that includes all transfer functions which belong to \mathcal{H}^{∞} and have degree not greater than m . Moreover, assume that

$$\alpha = \min_{g_0(z) \in \mathcal{H}_m^{\infty}} \max_{g(z) \in \mathcal{G}} \|W_{ai}(z)[g(z) - g_0(z)]\|_{\infty}, \quad (99)$$

$$g_0^*(z) = \arg \min_{g_0(z) \in \mathcal{H}_m^{\infty}} \max_{g(z) \in \mathcal{G}} \|W_{ai}(z)[g(z) - g_0(z)]\|_{\infty} \quad (100)$$

then, the desirable transfer function set \mathcal{G}_{opt} will be given as

$$\mathcal{G}_{opt} = \{g(z) \mid g(z) = g_0^*(z) + \alpha W_{ai}^{-1}(z)\Delta_a(z), \|\Delta_a(z)\|_{\infty} \leq 1\}. \quad (101)$$

On the other hand, when an upper bound of the \mathcal{H}^{∞} -norm of the plant transfer function and a series of the plant weakly corrupted time domain input-output data have been supplied, transfer function set \mathcal{G} can be parameterized as a linear fractional transformation and the cost function defined in (98) can be rewritten as

$$\begin{aligned} J(g_0(z)) &= \max_{g(z) \in \mathcal{G}} \|W_{ai}(z)[g(z) - g_0(z)]\|_{\infty} \\ &= \max_{\Delta(z) \in \mathcal{B}\mathcal{X}(z)} \|W_{ai}(z)\{F_{11}(z) + F_{12}(z)\Delta(z) \\ &\quad \cdot [I - F_{22}(z)\Delta(z)]^{-1}F_{21}(z) - g_0(z)\}\|_{\infty} \end{aligned}$$

$$\begin{aligned}
 &= \max_{\Delta(z) \in \mathcal{B}\mathcal{X}(z)} \left\| \mathcal{F}_l \left[\begin{pmatrix} W_{a1}(z)(F_{11}(z) - g_0(z)) & W_{a1}(z)F_{12}(z) \\ F_{21}(z) & F_{22}(z) \end{pmatrix} \right. \right. \\
 &\quad \left. \left. \cdot \Delta(z) \right] \right\|_{\infty}. \quad (102)
 \end{aligned}$$

Theorem 6: Assume that

$$\begin{aligned}
 &\min_{g_0(z) \in \mathcal{H}_m^{\infty}} \max_{\Delta(z) \in \mathcal{B}\mathcal{X}(z)} \left\| \mathcal{F}_l \left[\begin{pmatrix} W_{a1}(z)(F_{11}(z) - g_0(z)) & W_{a1}(z)F_{12}(z) \\ F_{21}(z) & F_{22}(z) \end{pmatrix} \right. \right. \\
 &\quad \left. \left. \cdot \Delta(z) \right] \right\|_{\infty} = \alpha
 \end{aligned}$$

and $g_0^*(z)$ is the solution of this minimization problem. Then

$$\mu_{\overline{\mathcal{X}}} \left[\begin{pmatrix} \frac{1}{\alpha} W_{a1}(F_{11} - g_0^*) & \frac{1}{\sqrt{\alpha}} W_{a1} F_{12} \\ \frac{1}{\sqrt{\alpha}} F_{21} & F_{22} \end{pmatrix} \right] = 1.$$

Proof: Since $\alpha = 0$ is a trivial case, we assume without loss of generality that $\alpha > 0$. From

$$\begin{aligned}
 &\min_{g_0(z) \in \mathcal{H}_m^{\infty}} \max_{\Delta(z) \in \mathcal{B}\mathcal{X}(z)} \left\| \mathcal{F}_l \left[\begin{pmatrix} W_{a1}(z)(F_{11}(z) - g_0(z)) & W_{a1}(z)F_{12}(z) \\ F_{21}(z) & F_{22}(z) \end{pmatrix} \right. \right. \\
 &\quad \left. \left. \cdot \Delta(z) \right] \right\|_{\infty} = \alpha \quad (103)
 \end{aligned}$$

and at $g_0^*(z)$ the cost function achieves its minimal value, we have

$$\begin{aligned}
 &\max_{\Delta(z) \in \mathcal{B}\mathcal{X}(z)} \left\| \mathcal{F}_l \left[\begin{pmatrix} \frac{1}{\alpha} W_{a1}(z)(F_{11}(z) - g_0^*(z)) & \frac{1}{\sqrt{\alpha}} W_{a1}(z)F_{12}(z) \\ \frac{1}{\sqrt{\alpha}} F_{21}(z) & F_{22}(z) \end{pmatrix} \right. \right. \\
 &\quad \left. \left. \cdot \Delta(z) \right] \right\|_{\infty} = 1. \quad (104)
 \end{aligned}$$

Since $\mu_{\underline{\mathcal{X}}}(F_{22}) < 1$ according to the proof of Theorem 5, from Lemma 1, we have

$$\mu_{\overline{\mathcal{X}}} \left[\begin{pmatrix} \frac{1}{\alpha} W_{a1}(F_{11} - g_0^*) & \frac{1}{\sqrt{\alpha}} W_{a1} F_{12} \\ \frac{1}{\sqrt{\alpha}} F_{21} & F_{22} \end{pmatrix} \right] = 1. \quad (105)$$

This completes the proof. ■

For the given transfer function matrix $F(z)$ and the frequency weight function $W_{a1}(z)$, define a scalar function $\overline{J}(\beta, g_0(z))$ as

$$\overline{J}(\beta, g_0(z)) = \mu_{\overline{\mathcal{V}}} \left[\begin{pmatrix} \frac{1}{\beta} W_{a1}(F_{11} - g_0) & \frac{1}{\sqrt{\beta}} W_{a1} F_{12} \\ \frac{1}{\sqrt{\beta}} F_{21} & F_{22} \end{pmatrix} \right] \quad (106)$$

for $\beta > 0$ and $g_0(z) \in \mathcal{H}_m^{\infty}$.

Note that since $\mu_{\underline{\mathcal{X}}}(F_{22}) < 1$, we have

$$\det \left(I - \begin{bmatrix} \frac{1}{\beta} W_{a1}(z)(F_{11}(z) - g_0(z)) & \frac{1}{\sqrt{\beta}} W_{a1}(z)F_{12}(z) \\ \frac{1}{\sqrt{\beta}} F_{21}(z) & F_{22}(z) \end{bmatrix} \cdot \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} \right)$$

$$\begin{aligned}
 &= \det(I - F_{22}(z)\Delta_2) \\
 &\quad \times \det \left(I - W_{a1}(z)[F_{11}(z) - g_0(z) \right. \\
 &\quad \left. + F_{12}(z)\Delta_2(I - F_{22}(z)\Delta_2)^{-1}F_{21}(z)] \left(\frac{1}{\beta} \Delta_1 \right) \right) \quad (107)
 \end{aligned}$$

at every $z \in \mathcal{D}$, for all $\Delta_1 \in \mathcal{C}$ and $\Delta_2 \in \mathcal{B}\mathcal{X}$. It is obvious that if $\beta_1 > \beta_2 > 0$, then

$$\overline{J}(\beta_1, g_0(z)) \leq \overline{J}(\beta_2, g_0(z)). \quad (108)$$

Therefore, when $\beta_1 > \beta_2 > 0$, we have

$$\min_{g_0(z) \in \mathcal{H}_m^{\infty}} \overline{J}(\beta_1, g_0(z)) \leq \min_{g_0(z) \in \mathcal{H}_m^{\infty}} \overline{J}(\beta_2, g_0(z)). \quad (109)$$

From the above discussion, it is apparent that the problem of obtaining the smallest transfer function set, which is determined by a fixed complexity nominal model and its additive error uncertainty bound, and includes all of the plant possible transfer functions deduced from the aforementioned plant information, can be transformed into a μ -synthesis problem, with combination of one free parameter search, i.e., to find the smallest β such that at this point the minimal structured singular value (with respect to $g_0(z) \in \mathcal{H}_m^{\infty}$) of the transfer function matrix

$$\begin{bmatrix} \frac{1}{\beta} W_{a1}(z)(F_{11}(z) - g_0(z)) & \frac{1}{\sqrt{\beta}} W_{a1}(z)F_{12}(z) \\ \frac{1}{\sqrt{\beta}} F_{21}(z) & F_{22}(z) \end{bmatrix}$$

according to the matrix structure $\overline{\mathcal{X}}$ is equal to 1. From the properties of the cost function $\overline{J}(\beta, g_0(z))$ defined in (106), i.e., $\overline{J}(\beta, g_0(z))$ will not increase with the increment of β , the algorithm for the search of the desirable β can be easily developed. A candidate procedure for this search may be as following.

Assume that two initial values of β , $0 < \beta_s^0 < \beta_l^0$ have been given which satisfy

$$\min_{g_0(z) \in \mathcal{H}_m^{\infty}} \overline{J}(\beta_s^0, g_0(z)) > 1, \quad \min_{g_0(z) \in \mathcal{H}_m^{\infty}} \overline{J}(\beta_l^0, g_0(z)) < 1.$$

In the N th iteration, let $\beta = \beta_s^{N-1} + (1 - \rho_N)\beta_l^{N-1}$. If $\min_{g_0(z) \in \mathcal{H}_m^{\infty}} \overline{J}(\beta, g_0(z)) > 1$, let $\beta_s^N = \beta$, $\beta_l^N = \beta_l^{N-1}$; otherwise, let $\beta_s^N = \beta_s^{N-1}$, $\beta_l^N = \beta$. If $0 < \rho_N < 1$ for all the iterations, it can be proved that this algorithm will converge to the desirable β . For detail, refer to [39].

In summary, we claim that the difficulty in developing the optimization algorithm for obtaining $g_0^*(z)$ and α that describe the desirable transfer function set \mathcal{G}_{opt} lies on μ -synthesis problem and the large dimension of the transfer function matrix $F(z)$.

V. CONCLUSION

In this paper, we investigated the structure of a transfer function set which includes all the possible plant transfer functions, under the situation that an upper bound of the \mathcal{H}^{∞} -norm of the plant transfer function is prescribed and the plant time domain input-output data are only weakly corrupted

in its output. Results show that such a transfer function set has an elegant structure, i.e., all of its element can be parameterized by a linear fractional transformation of a fixed transfer function matrix and a structure fixed, norm bounded, free transfer function matrix. The fixed transfer function matrix is completely determined by the plant measured input-output data, the noise level and the upper bound of the \mathcal{H}^∞ -norm of the plant transfer function. Moreover, it has been shown that the problem of obtaining the smallest transfer function set, which is defined by a fixed complexity nominal model and its additive error uncertainty bound and includes all of the plant possible transfer functions, can be transformed into a series of μ -synthesis problems.

While in this situation, the plant possible transfer function can be compactly expressed, the results in the last section suggest that it may be futile to pursue the analytic solution for the problem of searching the transfer function that belongs to \mathcal{H}^∞ , has a fixed complexity, and approximates all the plant possible transfer functions in the smallest additive error measured by a frequency weighted \mathcal{H}^∞ -norm. Moreover, although this problem can be converted to a series of μ -synthesis problems, it is still computationally difficult, due to the large dimension of the parameterization of the plant possible transfer function and the difficulties met in μ -synthesis problem itself. From these points, further research in developing computationally tractable approximation methods is still needed.

ACKNOWLEDGMENT

The authors would like to express their grateful acknowledgement to the anonymous referees for their suggestive comments and valuable criticisms.

REFERENCES

- [1] V. M. Adamjan, D. Z. Arov, and M. G. Krein, "Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem," *Math. Sbornik*, vol. 15, no. 1, pp. 31-73, 1971.
- [2] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis* (Translated by N. Kemmer). Edinburgh, Scotland: Oliver & Boyd, 1965.
- [3] E. W. Bai, "Adaptive quantification of model uncertainties by rational approximation," *IEEE Trans. Automat. Contr.*, vol. 36, no. 4, pp. 441-453, 1991.
- [4] J. C. Doyle, "An analysis of feedback systems with structured uncertainties," *IEE Proc.*, vol. 129, part D, pp. 242-250, 1982.
- [5] J. C. Doyle, K. Glover, P. Khargonekar, and B. A. Francis, "State-space solutions to standard \mathcal{H}_2 and \mathcal{H}_∞ control problems," *IEEE Trans. Automat. Contr.*, vol. 34, no. 8, pp. 831-847, 1989.
- [6] H. Dym and I. Gohberg, "Unitary interpolants, factorization indices and infinite Hankel block matrices," *J. Functional Anal.*, vol. 54, no. 3, pp. 229-289, 1983.
- [7] P. Eykhoff and P. C. Parks, Eds., *Special Issue on Identification and System Parameter Estimation*, *Automatica*, vol. 26, no. 1, 1990.
- [8] M. K. H. Fan, A. L. Tits, and J. C. Doyle, "Robustness in the presence of mixed parametric uncertainty and unmodeled dynamics," *IEEE Trans. Automat. Contr.*, vol. 36, no. 1, pp. 25-38, 1991.
- [9] M. Gevers, "Connecting identification and robust control," in *Proc. Workshop Modelling of Uncertainty in Control Systems*, Univ. California, Santa Barbara, 1992.
- [10] K. Glover, "All optimal Hankel-norm approximation of linear multivariable systems and their \mathcal{L}^∞ -error bounds," *Int. J. Contr.*, vol. 39, no. 6, pp. 1115-1193, 1984.
- [11] M. Green, K. Glover, D. Limebeer, and J. C. Doyle, "A J-spectral factorization approach to \mathcal{H}_∞ control," *SIAM J. Contr. Optimization*, vol. 28, no. 6, pp. 1350-1371, 1990.
- [12] G. C. Goodwin, M. Gevers, and B. Ninness, "Quantifying the error in estimated transfer functions with application to model order selection," *IEEE Trans. Automat. Contr.*, vol. 37, no. 7, pp. 913-928, 1992.
- [13] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*. New York: Chelsea, 1984.
- [14] G. X. Gu and P. P. Khargonekar, "A class of algorithms for identification in \mathcal{H}_∞ ," *Automatica*, vol. 28, no. 2, pp. 299-312, 1992.
- [15] A. J. Helmicki, C. A. Jacobson, and C. N. Nett, "Control oriented system identification: A worst-case/deterministic approach in \mathcal{H}_∞ ," *IEEE Trans. Automat. Contr.*, vol. 36, no. 10, pp. 1163-1176, 1991.
- [16] H. Kimura, "Robust stabilizability for a class of transfer functions," *IEEE Trans. Automat. Contr.*, vol. AC-29, no. 9, pp. 788-793, 1984.
- [17] ———, "On interpolation-minimization problem and system theory in hardy space," *Instrum. Contr.*, vol. 24, no. 7, pp. 23-32, 1985 (in Japanese).
- [18] ———, "From \mathcal{LQG} to \mathcal{H}^∞ ," *Instrum. Contr.*, vol. 29, no. 2, pp. 111-119, 1990 (in Japanese).
- [19] ———, "An essay on robust control," in *Proc. Workshop Modelling of Uncertainty in Control Systems*, Univ. California, Santa Barbara, 1992.
- [20] H. Kimura, Y. F. Lu, and R. Kawatani, "On the structure of \mathcal{H}^∞ control systems and related extensions," *IEEE Trans. Automat. Contr.*, vol. 36, no. 6, pp. 653-667, 1991.
- [21] R. L. Kosut, M. Lau, and S. Boyd, "Identification of systems with parametric and nonparametric uncertainty," in *Proc. Amer. Control Conf.*, San Diego, CA, 1990, pp. 2412-2417.
- [22] M. G. Krein and A. A. Nudel'man, "The Markov moment problem and extremal problems," *Translations Math. Monographs*, vol. 50, Amer. Math. Soc., Providence, RI, 1977.
- [23] S. Y. Kung and D. W. Lin, "Optimal Hankel-norm model reductions: Multivariable systems," *IEEE Trans. Automat. Contr.*, vol. AC-26, no. 4, pp. 832-852, 1981.
- [24] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
- [25] L. Ljung, B. Wahlberg, and H. Hjalmarsson, "Model Quality: The roles of prior knowledge and data information," in *Proc. IEEE Conf. Decision and Control*, Brighton, England, 1991, pp. 273-278.
- [26] P. J. Parker and R. Bitmead, "Adaptive frequency response identification," in *Proc. IEEE Conf. Decision and Control*, Los Angeles, CA, 1987, pp. 348-353.
- [27] A. Packard and J. C. Doyle, "The complex structured singular value," *Automatica*, vol. 29, no. 1, pp. 71-109, 1993.
- [28] K. Poolla, P. P. Khargonekar, A. Tikku, J. Krause, and K. Nagpal, "A time-domain approach to model validation," in *Proc. Amer. Control Conf.*, Chicago, IL, 1992, pp. 313-317.
- [29] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York: McGraw-Hill, 1986.
- [30] R. S. Smith and J. C. Doyle, "Towards a methodology for robust parameter identification," in *Proc. Amer. Control Conf.*, San Diego, CA, 1990, pp. 2395-2399.
- [31] J. F. Traub, G. W. Wasilkowski, and H. Wozniakowski, *Information-Based Complexity*. New York: Academic, 1988.
- [32] B. Wahlberg and L. Ljung, "Hard frequency-domain model error bounds from least-squares like identification techniques," *IEEE Trans. Automat. Contr.*, vol. 37, no. 7, pp. 900-912, 1992.
- [33] J. C. Willems, "Paradigms and puzzles in the theory of dynamical systems," *IEEE Trans. Automat. Contr.*, vol. 36, no. 3, pp. 259-294, 1991.
- [34] R. C. Younce and C. E. Rohrs, "Identification with nonparametric uncertainty," *IEEE Trans. Automat. Contr.*, vol. 37, no. 6, pp. 715-728, 1992.
- [35] G. Zames, "On the metric complexity of causal linear systems: ϵ -entropy and ϵ -dimension for continuous time," *IEEE Trans. Automat. Contr.*, vol. AC-24, no. 2, pp. 222-230, 1979.
- [36] T. Zhou and H. Kimura, "Minimal \mathcal{H}^∞ -norm of transfer functions consistent with prescribed finite input-output data," in *Proc. SICE '92*, Kumamoto, Japan, 1992, pp. 1079-1082.
- [37] ———, "Input-output extrapolation-minimization theorem and its applications to model validation and robust identification," in *Proc. Workshop Modelling of Uncertainty in Control Systems*, Univ. California, Santa Barbara, 1992.
- [38] ———, "Time domain identification for robust control," *Syst. Contr. Lett.*, vol. 20, no. 3, pp. 167-178, 1993.
- [39] ———, "Simultaneous identification of nominal model, parametric uncertainty and unstructured uncertainty for robust control," *Automatica*, vol. 30, no. 3, pp. 391-402, 1994.
- [40] ———, "Structure of model uncertainty for weakly corrupted plant," in *Proc. MTNS*, Regensburg, Germany, Aug. 1993.

- [41] T. Zhou, "Extrapolation, model validation and identification for robust control," Ph.D. dissertation, Dep. Mechanical Engineering for Computer-Controlled Machinery, Osaka Univ., Osaka, Japan, 1994.



Tong Zhou was born in Hunan Province, China, in 1964. He received the B.A. degree in control engineering from Chengdu Institute of Telecommunications Engineering in 1984, the M.S. degree in automatic control theory and application from the University of Electronic Science and Technology of China in 1988, the M.S. degree in electrical and computer engineering from Kanazawa University in 1991, and the Ph.D. degree in mechanical engineering for industrial machinery and systems from Osaka University in 1994.

From 1984 to 1986, he was a Research and Teaching Assistant at Chengdu Institute of Telecommunications Engineering. He joined the Department of Automation, University of Electronic Science and Technology of China in 1994, where he is presently an Associate Professor. His research interests include robust control theory, identification theory, adaptive control theory and their application to real world control engineering problems.



Hidenori Kimura (F'90) received the Bachelor, Master, and Doctor of Engineering degrees from the University of Tokyo in 1965, 1967, and 1970, respectively.

In 1970, he joined the Department of Control Engineering, Osaka University, where he was engaged in research and education of control engineering until 1987, when he moved to the Department of Mechanical Engineering for Computer-Controlled Machinery, Osaka University. In 1994, he joined the Department of Systems Engineering, Osaka University.

He is currently a Professor of Control Engineering at the Department of Mathematical Engineering and Information Physics, The University of Tokyo. His current research interests lie in H^∞ control, robust control, modeling for robust control and study of complex systems.

In 1972, 1983, and 1993, Dr. Kimura received the Paper Award from the SICE (the Society of Instrumentation and Control Engineers). Also, he was a recipient of the Paper Prize Award from IFAC in 1984 and 1990 and of the Outstanding Paper Award from the IEEE Control Systems Society in 1985. In 1990, he was awarded the Fellow of IELE. Since 1991, he has been a member of the Board of Governors of the IEEE Control Systems Society. He is the General Chair of the 35th CDC to be held in Japan in 1996.

Explicit Formulas for Optimally Robust Controllers for Delay Systems

Harry Dym, Tryphon T. Georgiou, and Malcolm C. Smith, *Member, IEEE*

Abstract—This paper considers single-input/single-output systems whose transfer functions take the form of a strictly proper rational function times a delay. A closed-form expression is presented for the controller which is optimally robust with respect to perturbations measured in the gap metric. The formula allows the \mathcal{H}_∞ loop-shaping procedure of Glover–McFarlane to be carried out explicitly for this class of systems without the need to first find a rational approximation of the plant. The form of the controller involves a certain algebra of “pseudo-derivation” operators. These operators, and their matrix generalizations, play a central role in the derivation of the controller. A discussion of the main properties of these operators will be given. An example will be presented of a controller design to achieve disturbance attenuation and robust set-point following for a plant with two lightly damped poles and a nontrivial time delay. The performance is compared, and shown to be superior, to that of a Smith predictor.

NOTATION

\mathbb{R} and \mathbb{C} denote the real and complex fields. $\mathcal{H}_2, \mathcal{H}_\infty$ denote the standard Hardy spaces in the right-half plane (RHP). \mathcal{L}_p denotes the standard Lebesgue space and $\|\cdot\|_p$ its norm. The inner product in \mathcal{L}_2 is denoted by $\langle \cdot, \cdot \rangle$. For $x \in \mathcal{L}_2(-\infty, \infty)$ we denote the Laplace transform of x by \hat{x} . If $G(s)$ is a matrix function of s , $G^*(s) := \overline{G(-\bar{s})}^T$ is the conjugate of $G(s)$. If $G(s)$ is real rational, $G^*(s) = G(-s)^T$. $\mathcal{H}_2^n, \mathcal{L}_2^n$, etc. denote corresponding spaces of vector-valued functions. $(\mathcal{H}_2^n)^\perp$ denotes the orthogonal complement of \mathcal{H}_2^n in $\mathcal{L}_2^n(-j\infty, j\infty)$. We denote by Π_+ (respectively, Π_-) the orthogonal projection from $\mathcal{L}_2^n(-j\infty, j\infty)$ onto \mathcal{H}_2^n (respectively, $(\mathcal{H}_2^n)^\perp$). Dimensions will be suppressed if they can be inferred from the context. The symbol $\text{spec}(A)$ designates the eigenvalues of A .

I. INTRODUCTION

IN this paper we consider the class of single-input/single-output delay systems of the form $P(s) = e^{-s\tau} P_0(s)$, where $P_0(s)$ is a strictly proper rational function and $\tau > 0$ is a time-delay. We present a closed form formula for controllers which are optimally robust with respect to gap/normalized coprime

factor uncertainty. This formula is a generalization of one given in [19] (see also [18]) for the case of a first-order $P_0(s)$. The approach also builds on the work of [24] which developed state-space formulas for computing optimal H_∞ -performance for such systems.

The derivation of the controllers is based on a synthesis of state-space techniques, associated with the rational part $P_0(s)$, along with the use of a certain commutative algebra of operators. These operators are defined as follows

$$\partial_\alpha f := \frac{f(s) - f(\alpha)}{s - \alpha}$$

(for $\alpha \in \mathbb{C}$ and $f(s)$ any meromorphic function which is analytic at α), and they represent an analog of the “backward shift” operator in the continuous-time setup. Some early use of these operators ([2] and [3]) was in the context of reproducing kernel Hilbert spaces and scattering theory—see [7] and [8] for more recent accounts, numerous references, and applications to interpolation theory. The formula for the optimal controllers contains infinite-dimensional “distributed delays” resulting from the action of the ∂_α ’s (or a suitable matrix generalization ∂_H) on the delay element $e^{-s\tau}$.

The controller formula derived in this paper allows the \mathcal{H}_∞ loop shaping design procedure of Glover–McFarlane to be carried out explicitly for the delay systems considered without the need to first find a rational approximation of the plant. We illustrate this fact by designing a controller to achieve disturbance attenuation and robust set-point following for a plant with two lightly damped poles and a nontrivial delay. This type of system is a suitable nominal model in a number of process control applications. We will show that our design has significantly better performance than a Smith predictor.

Research on the subject of \mathcal{H}_∞ control for distributed parameter systems has been on-going since the mid 1980’s. One of the objectives of this research has been to develop explicit formulas which provide insight into the limitations on performance imposed by infinite-dimensional elements such as delays. An additional goal has been to understand the effect of such elements on the structure of the controller, thereby giving an alternative design approach to finite-dimensional approximation of the plant. Early work was concerned with computing the optimal performance for problems such as weighted sensitivity minimization [10], [13]. Subsequent efforts focused mainly on developing computational methodologies for increasingly general situations [14], [22], [28], [29], [34] and for explicitly computing the controllers. The computation of the optimal performance involves finding the

Manuscript received December 1, 1993; revised May 25, 1994 and July 23, 1994. Recommended by Associate Editor, B. Lehman. This work was supported in part by the NSF, AFOSR, SERC, and the Nuffield Foundation.

H. Dym is with the Department of Theoretical Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel.

T. T. Georgiou is with the Department of Electrical Engineering, University of Minnesota, Minneapolis, MN 55455 USA.

M. C. Smith is with the Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K.

IEEE Log Number 9408279

maximum singular value (norm) of a certain generalized Hankel or Sarason operator. The construction of the optimal controller requires computing the associated singular vectors. Extra care is needed in case the infimum is not achieved in the optimization problem [11] or in case the underlying operator is not compact. (We mention incidentally that neither of these two cases arises in the problem treated here.) The computation of singular values and vectors usually exploits special structures in the family of systems considered, such as the presence of general \mathcal{H}_∞ inner factors multiplied by rational functions. In the case of delay systems, the computations were originally cast in the form of a certain two-point boundary value problem in the time domain [13], [24], [35] which in turn led to a solution in the form of a transcendental equation. Analogous computations can be carried out in the s -domain, or in the z -domain after a bilinear transformation, as in the skew-Toeplitz approach in [1] and [22]. The general methodology has been shown to apply to quite a general class of distributed parameter systems and provides a framework for obtaining a numerical solution [12], [23]. An alternative viewpoint for \mathcal{H}_∞ control of distributed parameter systems, based on an infinite dimensional state space approach, has been considered in [5] and [6] and the references therein.

The current work is part of a continuing effort to realize the original goals of the above research program. A formula for an \mathcal{H}_∞ controller is derived which takes a particularly simple form and is easily programmable. Moreover, the controller minimizes a certain cost functional which has proved to be especially convenient for design. We would like to mention that, in some recent work [30], an algorithm has been outlined for the computation of suboptimal controllers of delay systems for robustness in the gap. At the moment, it is an interesting open problem to seek an explicit formula of the type given in this paper for the suboptimal case.

The paper is organized as follows. In Section II some background material is reviewed on the properties and computation of gap/normalized coprime factor optimally robust controllers. In Section III the singular value/vector equations of Partington–Glover, and the procedure to find the gap optimal controller from them, are summarized. In Section IV the (scalar) operator ∂_α is introduced, and the main properties of the operator, which will be used in the controller derivation, will be described. In Section V the general procedure for the computation of the optimal controller will be carried out for the case where $P(s) = e^{-s\tau}/(s - \sigma)$. The purpose of this section is to motivate the computational steps that are necessary in the higher order case. In Section VI a generalization of the ∂_α operator to the matrix case will be presented together with its properties. In Section VII the general controller formula is derived and, in Section VIII, an illustrative example is presented.

II. BACKGROUND

We consider the feedback configuration of Fig. 1, denoted by $[P, K]$, where P and K are the transfer functions of the plant and the controller, respectively (and are ratios of \mathcal{H}_∞ -functions). We say that $[P, K]$ is stable if the closed-loop

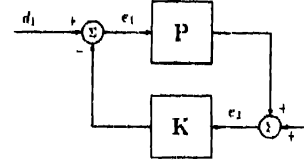


Fig. 1 Standard feedback configuration.

transfer functions, mapping d_i to e_j ($i, j = 1, 2$), belong to \mathcal{H}_∞ . If $[P, K]$ is stable then the following quantity can be defined

$$b_{P,K} := (I + KP)^{-1}(I, K)$$

which is the gap metric robustness radius for the feedback system. That is, $b_{P,K}$ is the radius of the largest ball of uncertainty about P , measured in the gap metric, which is stabilized by K . For the definition of the gap metric, connections with normalized coprime factorizations, and associated results on robust stabilization the reader is referred to [16] and [19].

It appears that $b_{P,K}$ is a useful quantity to maximize for a feedback system, especially if it is weighted appropriately. In fact, the maximization of $b_{WP,K}$, for some weighting function W chosen by the designer, is the basis of the Glover–McFarlane \mathcal{H}_∞ loop shaping method [21]. This problem is also the same as the optimal robustness problem in the weighted gap metric (see [4] and [15]). From the point of view of computations, it is sufficient to consider the unweighted case only. This will be our point of view until the example in the last section.

We now summarize some background results on the optimization of $b_{P,K}$. Let $P = N/M$, where $M, N \in \mathcal{H}_\infty$, is a normalized coprime factorization of P (i.e., $NN^* + MM^* = 1$) and let $F = (-N, M)$. We define the optimal robustness radius

$$b_{\text{opt}}(P) := \sup_{K \text{ stabiliz}} b_{P,K}. \quad (1)$$

It can be shown that the following formula holds

$$b_{\text{opt}}(P) = (1 - \|F\|^2)^{1/2} \quad (2)$$

where F is the Hankel operator defined by $F = H_- F^*|_{\mathcal{H}_2}$ (see [19] and [16]).

If F is continuous on the compactified right-half plane then it follows (from Hartman's Theorem) that the operator F is compact and hence that it achieves its norm $\|F\|$, which is equal to the maximal singular value of the operator. Moreover if $0 < \lambda = \|F\| < 1$ then there is a unique optimally robust controller which can be determined from λ and any corresponding singular vector. Specifically, let $\hat{x} \in \mathcal{H}_2$, $\hat{y} = (\hat{y}_1, \hat{y}_2) \in \mathcal{H}_2^{1 \times 2}$ satisfy

$$F\hat{x} = \lambda\hat{y}^*, \quad (3)$$

$$F^*\hat{y}^* = \lambda\hat{x} \quad (4)$$

and define $\hat{u} = H_+ F^* \lambda \hat{x}$. Then the optimal compensator has a transfer function $K = -\hat{u}_1/\hat{u}_2$. See [18] for a more detailed explanation of these facts and further references.

III. COMPUTATION OF SINGULAR VALUES AND VECTORS

In this section we consider a single-input/single-output system with transfer function of the form $P(s) = e^{-s\tau} P_0(s)$ where $P_0(s)$ is a strictly proper rational function and τ is a time-delay. Since $e^{-s\tau}$ is inner we can find the normalized coprime factors of $P(s)$ by finding the normalized coprime factors of $P_0(s)$ using the state-space construction of [20] and [31] and then multiplying the numerator by $e^{-s\tau}$. Let $P_0(s) = C(sI - A_0)^{-1}B$ be a minimal realization with state dimension n . We define R to be the stabilizing solution of the Riccati equation

$$A_0 R + R A_0 - R C^* C R + B B^* = 0 \quad (5)$$

and set $A = A_0 + EC$ where $E = -RC^*$. Then $\text{spec}(A)$ is in the open left-half plane, and

$$F = (-N, M) = (0, I) + C(sI - A)^{-1}(-B e^{-s\tau}, E). \quad (6)$$

We assume that $P(s) \not\equiv 0$ which implies that $\|F\| > 0$. Since $F(s)$ is right invertible over \mathcal{H}_∞ , which is equivalent to P being stabilizable, $\|F\| < 1$ (see [16, Sections III and VII]). Finally we note that, since $F(s)$ is continuous on the compactified right-half plane, the optimal controller for $P(s)$ can be found from the singular values and vectors of F .

We now summarize the formulas of Partington and Glover [24] for the computation of the singular values of F . The idea is to give a time-domain realization of (3) and (4). These are the equations

$$\begin{aligned} w(\tau) &= \int_{\tau}^{\infty} e^{A^*(u-\tau)} C^* x(u) du, \\ \dot{w}(t) &= -A^* w(t) - C^* x(t) \quad \text{for } 0 \leq t \leq \tau, \\ \lambda y_1(t) &= -B^* w(\tau - t) \quad \text{for } 0 \leq t \leq \tau, \\ \lambda \begin{bmatrix} y_1(t + \tau) \\ y_2(t) \end{bmatrix} &= \begin{bmatrix} -B^* \\ E^* \end{bmatrix} e^{A^* t} w(0) \quad \text{for } t > 0 \end{aligned} \quad (7)$$

and

$$v(0) = \int_0^{\infty} e^{A u} (-B y_1(u + \tau) + E y_2(u)) du, \quad (8)$$

$$\begin{aligned} \dot{v}(t) &= A v(t) - B y_1(\tau - t) \quad \text{for } 0 \leq t \leq \tau, \\ \lambda x(t) &= \begin{cases} C v(t) & \text{for } 0 \leq t \leq \tau \\ C e^{A(t-\tau)} v(\tau) & \text{for } t \geq \tau \end{cases} \end{aligned} \quad (9)$$

These equations can be solved by considering a two-point boundary value problem on the interval $[0, \tau]$. On this interval the differential equations become

$$\frac{d}{dt} \begin{bmatrix} v \\ w \end{bmatrix} = H \begin{bmatrix} v \\ w \end{bmatrix} \quad (10)$$

where

$$H = \begin{bmatrix} A & \lambda^{-1} B B^* \\ -\lambda^{-1} C^* C & -A^* \end{bmatrix}. \quad (11)$$

Thus

$$\begin{bmatrix} v(t) \\ w(t) \end{bmatrix} = e^{Ht} \begin{bmatrix} v(0) \\ w(0) \end{bmatrix}. \quad (12)$$

Next, by substituting for x and y in (8) and (7), we obtain the relations

$$v(0) = \lambda^{-1} R w(0). \quad (13)$$

$$w(\tau) = \lambda^{-1} S v(\tau) \quad (14)$$

where S is the unique solution of the Lyapunov equation

$$A^* S + S A + C^* C = 0. \quad (15)$$

(Incidentally, observe that R satisfies the Lyapunov equation $AR + RA^* + BB^* + EE^* = 0$.) This leads to the equation

$$0 = L(\lambda) w(0)$$

where

$$L(\lambda) := \left\{ \begin{bmatrix} -S & \lambda I \end{bmatrix} \exp(H\tau) \begin{bmatrix} R \\ \lambda I \end{bmatrix} \right\}. \quad (16)$$

Proposition 1 [24]: A necessary and sufficient condition for λ to be a singular value of $F = \Pi F^*|_{\mathcal{H}_2}$ is that $\det L(\lambda) = 0$. \square

It follows that, to find the optimal controller for $P(s)$, we can use the following procedure:

- P1) Find the largest λ in $[0, 1]$ such that $\det L(\lambda) = 0$ and then find a nonzero vector w_0 such that $L(\lambda)w_0 = 0$. (By the assumptions on $P(s)$, $\lambda \neq 0, 1$.) Then find $v_0 = \lambda^{-1} R w_0$ and solve for $v(t)$ on $[0, \tau]$ using (12) with $v(0) = v_0$ and $w(0) = w_0$.
- P2) Calculate $\lambda x(t)$ on $[0, \infty)$ from (9).
- P3) Find $u = \Pi_+ K^* \lambda \hat{x}$, whence $K(s) = -u_1/\hat{u}_2$ [18].

IV. THE OPERATOR ∂_α

Let $f(s)$ be analytic in $\Omega \subset \mathbb{C}$, and define

$$(\partial_\alpha f)(s) := \frac{f(s) - f(\alpha)}{s - \alpha}$$

for $\alpha \in \Omega$. Then ∂_α defines a linear operator and has a number of interesting properties (see, e.g., [7] and [8]). Several of these properties will be exploited in this paper and we list these below. Note that ∂_α resembles a derivative, but without a limiting process. With $\alpha = 0$ the operator is the adjoint of the shift operator on \mathcal{H}_2 of the disc. Thus, ∂_α can be thought of as a generalized backward shift. It is readily verified that it satisfies the resolvent identity

$$\partial_\alpha - \partial_\beta = (\alpha - \beta) \partial_\alpha \partial_\beta \quad (17)$$

for any pair of points $\alpha, \beta \in \mathbb{C}$ which are points of analyticity for the class of functions acted upon. A direct consequence is that $\partial_\alpha \partial_\beta = \partial_\beta \partial_\alpha$. The following pseudo-derivation identities take the place of a product rule

$$(\partial_\alpha f g)(s) = (\partial_\alpha f)(s) g(s) + f(\alpha) (\partial_\alpha g)(s), \quad (18)$$

$$(\partial_\alpha f g)(s) = f(s) (\partial_\alpha g)(s) + (\partial_\alpha f)(s) g(\alpha). \quad (19)$$

(They differ from the usual product rule of differentiation by the fact that one of the two terms involves evaluation of one of the functions at the point α .) If $\alpha + \gamma \neq 0$ then

$$\partial_\alpha \frac{1}{s + \gamma} = \frac{-1/(\alpha + \gamma)}{(s + \gamma)}. \quad (20)$$

If $\alpha + \gamma \neq 0$ and g is analytic at α , then (18), (19), and (20) give the following two identities

$$\partial_\alpha \frac{g}{s + \gamma} = \frac{1}{s + \gamma} \left(\partial_\alpha g - \frac{1}{\alpha + \gamma} g(\alpha) \right), \quad (21)$$

$$\partial_\alpha \frac{g}{s + \gamma} = \frac{1}{\alpha + \gamma} \left(\partial_\alpha - \frac{1}{s + \gamma} \right) g. \quad (22)$$

If $\operatorname{Re}(\gamma) > 0$, and $g \in \mathcal{H}_2$ then

$$\Pi_+ \frac{1}{s - \gamma} g = \partial_\gamma g. \quad (23)$$

Finally we present the following result which will be needed in the controller derivations of Sections V and VII.

Proposition 2: Let $\operatorname{Re}(\gamma) > 0$ and $\alpha + \gamma \neq 0$. Suppose $F \in \mathcal{H}_\infty$ admits a representation of the form $F(s) = c + \int_0^\infty e^{-st} f(t) dt$ with c constant and $f(t) \in \mathcal{L}_1[0, \infty)$. Then $\Pi_+ F^* \partial_\alpha e^{-s\tau} / (s + \gamma)^k$ is analytic in $\mathbb{C} \setminus \{-\gamma\}$, and

$$\Pi_+ F^* \partial_\alpha \frac{1}{(s + \gamma)^k} = \partial_\alpha \Pi_+ F^* \frac{1}{(s + \gamma)^k} \quad (24)$$

for $k = 1, 2, \dots$.

Proof: We consider first the case $k = 1$. Under the given assumptions it is readily checked that

$$\begin{aligned} \Pi_+ F^* \partial_\alpha \frac{1}{s + \gamma} &= -\frac{1}{\alpha + \gamma} \Pi_+ F^* \frac{1}{s + \gamma} \\ &\quad - \frac{F^*(-\gamma)}{\alpha + \gamma} \frac{1}{s + \gamma} \end{aligned} \quad (25)$$

$$\begin{aligned} &= \partial_\alpha \frac{F^*(-\gamma)}{s + \gamma} \\ &= \partial_\alpha \Pi_+ F^* \frac{1}{s + \gamma}. \end{aligned} \quad (26)$$

Equation (25) shows that $\Pi_+ F^* \partial_\alpha 1 / (s + \gamma)$ is analytic in $\mathbb{C} \setminus \{-\gamma\}$. The proof now proceeds in two steps. We first introduce the natural projection operator, in the Wiener algebra on the line, defined as follows

$$P : \int_{-\infty}^\infty e^{-st} g(t) dt \rightarrow \int_0^\infty e^{-st} g(t) dt$$

with $g \in \mathcal{L}_1(\mathbb{R})$ and $s \in j\mathbb{R}$.

Step 1: Let $F_1(s) = F(s) - c = \int_0^\infty e^{-st} f(t) dt$. Then $PF_1^* e^{-s\tau}$ is an entire function and

$$PF_1^* \partial_\alpha e^{-s\tau} = \partial_\alpha PF_1^* e^{-s\tau}$$

for every choice of $\alpha \in \mathbb{C}$.

Proof of Step 1: The proof is an adaptation of [9, Lemma 4.4]. Observe that

$$\begin{aligned} PF_1^* e^{-s\tau} &= P \int_{-\infty}^0 e^{-st} f^*(t) dt \cdot e^{-s\tau} \\ &= P \int_{-\infty}^\tau e^{-su} f^*(u - \tau) du \\ &= \int_0^\tau e^{-su} f^*(u - \tau) du \end{aligned}$$

and hence $PF_1^* e^{-s\tau}$ is entire since it is the Laplace transform of a function with support on the interval $[0, \tau]$. Further

$$\begin{aligned} \partial_\alpha PF_1^* e^{-s\tau} &= \int_0^\tau (\partial_\alpha e^{-su}) f^*(u - \tau) du \\ &= \int_0^\tau \int_0^u e^{-(s-\alpha)v} dv (-e^{\alpha u}) f^*(u - \tau) du \\ &= - \int_0^\tau \int_0^\tau e^{\alpha(v-u)} f^*(u - \tau) du dv \\ &= - \int_0^\tau \int_v^\tau e^{\alpha(x-\tau)} f^*(v - x) dx dv. \end{aligned} \quad (27)$$

On the other hand

$$\begin{aligned} PF_1^* \partial_\alpha e^{-s\tau} &= P \int_0^\tau e^{-sx} (-e^{\alpha(x-\tau)}) dx \int_{-\infty}^0 e^{-st} f^*(t) dt \\ &= P \int_0^\tau (-e^{\alpha(x-\tau)}) \int_{-\infty}^x e^{-sv} f^*(v - x) dv dx \end{aligned}$$

which turns into expression (27) after interchanging the order of integration and then taking the projection. ∇

Step 2: The function $\Pi_+ F^* \partial_{-\gamma} e^{-s\tau}$ is an entire function, and

$$\Pi_+ F^* \partial_\alpha \partial_{-\gamma} e^{-s\tau} = \partial_\alpha \Pi_+ F^* \partial_{-\gamma} e^{-s\tau}$$

for every complex number α such that $\alpha + \gamma \neq 0$.

Proof of Step 2: It follows using Step 1 that $\Pi_+ F_1^* \partial_{-\gamma} e^{-s\tau}$ is entire. Clearly $\Pi_+ c^* \partial_{-\gamma} e^{-s\tau} = c^* \partial_{-\gamma} e^{-s\tau}$ is also entire. We now verify the formula for each of the components of $F(s) = c + F_1(s)$ separately. The first is easy

$$\Pi_+ c^* \partial_\alpha \partial_{-\gamma} e^{-s\tau} = c^* \partial_\alpha \partial_{-\gamma} e^{-s\tau} = \partial_\alpha \Pi_+ c^* \partial_{-\gamma} e^{-s\tau}.$$

Next, for $\alpha \neq -\gamma$ we have

$$\begin{aligned} \Pi_+ F_1^* \partial_\alpha \partial_{-\gamma} e^{-s\tau} &= \Pi_+ F_1^* \frac{\partial_\alpha - \partial_{-\gamma}}{\alpha + \gamma} e^{-s\tau} \\ &= PF_1^* \frac{\partial_\alpha - \partial_{-\gamma}}{\alpha + \gamma} e^{-s\tau} \\ &= \frac{\partial_\alpha - \partial_{-\gamma}}{\alpha + \gamma} PF_1^* e^{-s\tau} \\ &= \partial_\alpha \partial_{-\gamma} PF_1^* e^{-s\tau} \\ &= \partial_\alpha PF_1^* \partial_{-\gamma} e^{-s\tau} \\ &= \partial_\alpha \Pi_+ F_1^* \partial_{-\gamma} e^{-s\tau} \end{aligned}$$

using Step 1. The operator P is needed here because $F_1^* e^{-s\tau}$ may not belong to $\mathcal{L}_2(j\mathbb{R})$ and hence Π_+ cannot always be applied to it directly. ∇

With the help of the preceding analysis, we can now prove the proposition for the case $k = 1$. For $\alpha \neq -\gamma$

$$\begin{aligned} \Pi_+ F^* \partial_\alpha \frac{e^{-s\tau}}{s + \gamma} &= \Pi_+ F^* \partial_\alpha \left\{ \partial_{-\gamma} e^{-s\tau} + \frac{e^{\gamma\tau}}{s + \gamma} \right\} \\ &= \partial_\alpha \Pi_+ F^* \left\{ \partial_{-\gamma} e^{-s\tau} + \frac{e^{\gamma\tau}}{s + \gamma} \right\} \end{aligned}$$

using (26) and Step 2. This establishes the required equality.

To complete the proof we note that the result for the case of $k = 1$ can be applied to

$$g_n(s) = \frac{1}{(s + \gamma) \left(s + \gamma + \frac{1}{n}\right) \cdots \left(s + \gamma + \frac{k-1}{n}\right)}$$

for n sufficiently large, to give

$$\Pi_+ F^* \partial_\alpha e^{-s\tau} g_n(s) = \partial_\alpha \Pi_+ F^* e^{-s\tau} g_n(s).$$

By letting $n \rightarrow \infty$ we can check in turn that

$$g_n(s) \rightarrow g(s) = \frac{1}{(s + \gamma)^k} \text{ in } \mathcal{L}_2(j\mathbb{R}),$$

$$\partial_\alpha g_n(s) \rightarrow \partial_\alpha g(s) \text{ in } \mathcal{L}_2(j\mathbb{R}), \text{ and}$$

$$\Pi_+ F^* e^{-s\tau} g_n(s) \rightarrow \Pi_+ F^* e^{-s\tau} g(s)$$

pointwise in $\mathbb{C} \setminus \{-\gamma\}$. Thus

$$\begin{aligned} \Pi_+ F^* \partial_\alpha e^{-s\tau} g(s) &= \lim_{n \rightarrow \infty} \Pi_+ F^* \partial_\alpha e^{-s\tau} g_n(s) \\ &= \lim_{n \rightarrow \infty} \partial_\alpha \Pi_+ F^* e^{-s\tau} g_n(s) \\ &= \partial_\alpha \Pi_+ F^* e^{-s\tau} g(s). \end{aligned} \quad \square$$

V. FIRST-ORDER CASE

In the present section we consider the simplest example of a plant having transfer function of the form "time-delay \times rational function." $P(s) = e^{-s\tau}/(s - \sigma)$. An expression for the optimal controller was derived in [18]. A simplified formula involving the operator ∂_α was given in [19]. Below we present a new derivation of the optimal controller making explicit use of certain properties of the operator ∂_α . The computations illustrate several key steps which motivate analogous steps in the general case.

Proposition 3: Let $P(s) = e^{-s\tau}/(s - \sigma)$ where σ is any real constant and $\tau > 0$.

- a) The optimal robustness radius for gap ball uncertainty is given by

$$b_{\text{opt}}(P) = (1 - \lambda_{\max}^2)^{1/2} \quad (28)$$

where λ_{\max} is the largest solution of the following transcendental equation

$$e^{2\mu\tau} = \frac{\sigma + \mu}{\sigma - \mu} \left(\frac{\gamma - \mu}{\gamma + \mu} \right)^2 \quad (29)$$

where $\mu = \mu(\lambda) = \sqrt{\gamma^2 - \lambda^{-2}}$ and $\gamma = \sqrt{1 + \sigma^2}$.

- b) The transfer function of the optimal controller is given by

$$K_{\text{opt}}(s) = \frac{1}{\sqrt{\sigma^2 - \mu^2} - \frac{\gamma^2 - \mu^2}{2\mu} ((\sigma + \mu)\partial_\mu - (\sigma - \mu)\partial_{-\mu}) e^{-s\tau}} \quad (30)$$

where $\mu = \sqrt{\gamma^2 - \lambda_{\max}^{-2}}$, in the case $\mu \neq 0$. (A general formula covering also the case $\mu = 0$ is given in Theorem 1.)

Proof: Setting $B = C = 1$, $A_0 = \sigma$ gives $A = -\gamma$, $R = \sigma + \gamma$ and $S = 1/2\gamma$, whence

$$F(s) = \left(-\frac{e^{-s\tau}}{s + \gamma}, \frac{s - \sigma}{s + \gamma} \right).$$

Then

$$H = \begin{pmatrix} -\gamma & \lambda^{-1} \\ -\lambda^{-1} & \gamma \end{pmatrix}$$

with eigenvalues $\pm\mu$ where $\mu = \sqrt{\gamma^2 - \lambda^{-2}}$. Note that μ can be complex but $\pm\mu \neq \gamma$. An eigenvalue-eigenvector decomposition of H leads to

$$\begin{aligned} e^{Ht} &= \begin{pmatrix} \lambda^{-1} & \lambda^{-1} \\ \gamma + \mu & \gamma - \mu \end{pmatrix} \begin{pmatrix} e^{\mu t} & 0 \\ 0 & e^{-\mu t} \end{pmatrix} \\ &\quad \cdot \begin{pmatrix} \gamma - \mu & -\lambda^{-1} \\ -(\gamma + \mu) & \lambda^{-1} \end{pmatrix} \frac{-\lambda}{2\mu}. \end{aligned}$$

By substituting into (16) it follows that

$$\begin{aligned} L(\lambda) &= 0 \Leftrightarrow e^{2\mu\tau} \\ &= \frac{-\lambda^{-2} + 2\gamma(\gamma - \mu)}{-\lambda^{-2} + 2\gamma(\gamma + \mu)} \cdot \frac{(\sigma + \gamma)(\gamma + \mu) - 1}{(\sigma + \gamma)(\gamma - \mu) - 1} \\ &= \frac{(\gamma - \mu)^2}{(\gamma + \mu)^2} \cdot \frac{(\sigma + \gamma)(\sigma + \mu)}{(\sigma + \gamma)(\sigma - \mu)}. \end{aligned}$$

This gives (29). Also (28) follows from (2). Setting $w_0 = 1$ gives $v_0 = \lambda^{-1}R$, which leads to

$$\begin{aligned} v(t) &= (1 \ 0) e^{Ht} \begin{pmatrix} \lambda^{-1} R \\ 1 \end{pmatrix} \\ &= \alpha e^{\mu t} + \beta e^{-\mu t} \end{aligned}$$

on $[0, \tau]$, where

$$\begin{aligned} \alpha &= \frac{-(\sigma + \gamma)(\sigma - \mu)}{2\mu\lambda}, \\ \beta &= \frac{(\sigma + \gamma)(\sigma + \mu)}{2\mu\lambda}. \end{aligned}$$

(The calculation of α and β is eased by judicious use of the fact that $1 = \gamma^2 - \sigma^2$.) Thus

$$\lambda r(t) = \begin{cases} \alpha e^{\mu t} + \beta e^{-\mu t} & \text{on } [0, \tau] \\ e^{-\gamma(t-\tau)} (\alpha e^{\mu\tau} + \beta e^{-\mu\tau}) & \text{on } [\tau, \infty). \end{cases}$$

Hence

$$\begin{aligned} \lambda \dot{r} &= \int_0^\infty \lambda r(t) e^{-st} dt \\ &= -\alpha e^{\mu\tau} \left(\partial_\mu - \frac{1}{s + \gamma} \right) e^{-s\tau} \\ &\quad - \beta e^{-\mu\tau} \left(\partial_{-\mu} - \frac{1}{s + \gamma} \right) e^{-s\tau} \\ &= (-\alpha(\gamma + \mu) e^{\mu\tau} \partial_\mu - \beta(\gamma - \mu) e^{-\mu\tau} \partial_{-\mu}) \frac{e^{-s\tau}}{s + \gamma} \end{aligned}$$

where the first step uses direct integration and the second step follows from (22). Next

$$\begin{aligned}\Pi_+ F^* \lambda \hat{x} &= \Pi_+ \left(\frac{e^{s\tau}/(s-\gamma)}{(s+\sigma)/(s-\gamma)} \right) \lambda x \\ &= (-\alpha(\gamma+\mu)e^{\mu\tau}\partial_\mu - \beta(\gamma-\mu)e^{-\mu\tau}\partial_{-\mu}) \\ &\quad \cdot \Pi_+ \left(\frac{e^{s\tau}/(s-\gamma)}{(s+\sigma)/(s-\gamma)} \right) \frac{e^{-s\tau}}{s+\gamma} \\ &= (-\alpha(\gamma+\mu)e^{\mu\tau}\partial_\mu - \beta(\gamma-\mu)e^{-\mu\tau}\partial_{-\mu}) \\ &\quad \cdot \left(\frac{-1/2\gamma(s+\gamma)}{(1+(\sigma+\gamma)\partial_\gamma)e^{-s\tau}/(s+\gamma)} \right),\end{aligned}$$

where we have used Proposition 2 in the first step (which is applicable because $\pm\mu + \gamma \neq 0$), and (20) and (23) in the second step. Using (20), the first element of $\Pi_+ F^* \lambda \hat{x}$ becomes

$$\frac{-1}{2\gamma(s+\gamma)}(\alpha e^{\mu\tau} + \beta e^{-\mu\tau}) = \frac{-(\sigma+\gamma)\sqrt{\sigma^2}}{\lambda(\gamma^2 - \mu^2)(s+\gamma)}$$

where the equality follows by substituting for α, β and $e^{\mu\tau} = \sqrt{(\sigma+\mu)/(\sigma-\mu)(\gamma-\mu)/(\gamma+\mu)}$. The second element of $\Pi_+ F^* \lambda \hat{x}$ is equal to

$$\begin{aligned}&\left\{ (-\alpha(\gamma+\mu)e^{\mu\tau}) \left(1 + \frac{\sigma+\gamma}{\mu-\gamma} \right) \partial_\mu \right. \\ &\quad + (-\beta(\gamma-\mu)e^{-\mu\tau}) \left(1 + \frac{\sigma+\gamma}{-\mu-\gamma} \right) \partial_{-\mu} + (\sigma+\gamma) \\ &\quad \left. \frac{\alpha(\gamma+\mu)}{\mu-\gamma} e^{\mu\tau} + \frac{\beta(\gamma-\mu)}{-\mu-\gamma} e^{-\mu\tau} \right\} \partial_\gamma \frac{e^{-s\tau}}{s+\gamma}\end{aligned}$$

after using (17). In fact, by substituting for α, β and $e^{\mu\tau}$, the coefficient of ∂_γ in Δ is easily seen to be zero. Next, using similar substitutions, we get

$$\begin{aligned}\Delta &= \frac{2\mu\lambda}{(\sigma+\gamma)\sqrt{\sigma^2 - \mu^2}} \{ -(\sigma+\mu)\partial_\mu + (\sigma-\mu)\partial_{-\mu} \} s + \gamma \\ &\quad \frac{2\mu\lambda(s+\gamma)}{(\sigma+\gamma)\sqrt{\sigma^2 - \mu^2}} \{ -(\sigma+\mu)\partial_\mu + (\sigma-\mu)\partial_{-\mu} \} e^{-s\tau} \\ &\quad + \frac{(\sigma+\gamma)\sqrt{\sigma^2 - \mu^2}}{2\mu\lambda(s+\gamma)} \frac{\sqrt{\sigma^2}}{\gamma^2 - \mu^2} 2\mu\end{aligned}$$

where the second step uses (21). Taking the quotient of (31) and the above expression for Δ gives the required result. \square

VI. THE MATRIX ∂ -OPERATOR

To facilitate the controller computations in higher order cases we will need a matrix generalization of the operator ∂_α and its properties.

Let $A \in \mathbb{C}^{n \times n}$ and let $F(s)$ be an $n \times r$ matrix whose elements are analytic in an open simply connected region $\Omega \subset \mathbb{C}$ for which $\text{spec}(A) \subset \Omega$. Then we define

$$(\partial_A F)(s) = \frac{1}{2\pi j} \int_{\gamma_1} \frac{1}{z-s} (zI - A)^{-1} F(z) dz \quad (32)$$

for s inside γ_1 , where γ_1 is any simple closed curve (with positive orientation) inside Ω which encloses $\text{spec}(A)$. We also define $(\partial_A F)(s)$ by the same formula if either A or $F(s)$ is scalar. The above definition (32) is independent of the choice

of γ_1 and can be evaluated explicitly in terms of the Jordan decomposition of A and the values of F and its derivatives at the points of the spectrum of A , as in (39) below. The present contour integral formulation is, however, often easier to work with. In particular, it follows from Cauchy's theorem that (32) yields

$$(\partial_\alpha F)(s) = \frac{F(s) - F(\alpha)}{s - \alpha}$$

for $s, \alpha \in \Omega$, when $A = \alpha$ is a scalar. This is consistent with the definition given in Section IV. On the other hand, if F is equal to a scalar function f and A is either a matrix or a scalar, it is standard to define

$$f(A) = \frac{1}{2\pi j} \int_{\gamma_1} (zI - A)^{-1} f(z) dz$$

where γ_1 is any simple closed curve inside Ω which encloses $\text{spec}(A)$ (see e.g., [26, Chapter VI.3]). Using the identity

$$(zI - sI)^{-1} - (zI - A)^{-1} = (zI - A)^{-1} (sI - A) (zI - sI)^{-1}$$

we see that

$$\begin{aligned}(\partial_A f)(s) &= (sI - A)^{-1} \left\{ \frac{1}{2\pi j} \int_{\gamma_1} \frac{1}{z-s} f(z) dz I \right. \\ &\quad \left. - \frac{1}{2\pi j} \int_{\gamma_1} (zI - A)^{-1} f(z) dz \right\} \\ &= (sI - A)^{-1} (f(s)I - f(A)).\end{aligned}$$

Let $A \in \mathbb{C}^{n \times n}$ and let $F(s)$ be an $n \times r$ matrix whose elements are analytic in an open simply connected region $\Omega \subset \mathbb{C}$. Take $\alpha \in \Omega$ with $\alpha \notin \text{spec}(A)$. Then we can show that the following generalizations of (20)–(22) hold

$$\begin{aligned}\partial_\alpha (sI - A)^{-1} &= -(sI - A)^{-1} (\alpha I - A)^{-1} \\ &= -(\alpha I - A)^{-1} (sI - A)^{-1}, \quad (33)\end{aligned}$$

$$\begin{aligned}\partial_\alpha ((sI - A)^{-1} F(s)) &= (\alpha I - A)^{-1} (\partial_\alpha F(s)) \\ &\quad - (sI - A)^{-1} F(s), \quad (34)\end{aligned}$$

$$\begin{aligned}\partial_\alpha ((sI - A)^{-1} F(s)) &= (sI - A)^{-1} (\partial_\alpha F(s)) \\ &\quad - (\alpha I - A)^{-1} F(\alpha). \quad (35)\end{aligned}$$

For the matrix ∂ operator, a version of the resolvent identity still holds. Let $A, B \in \mathbb{C}^{n \times n}$ with $AB = BA$ then

$$\partial_A - \partial_B = (A - B) \partial_A \partial_B \quad (36)$$

where the operators act on a matrix $F(s)$ whose elements are analytic in an open simply connected region $\Omega \subset \mathbb{C}$ containing $\text{spec}(A) \cup \text{spec}(B)$. To see this, take two simple closed curves γ_1, γ_2 inside Ω , both containing $\text{spec}(A) \cup \text{spec}(B)$, and with

γ_1 strictly inside γ_2 . Then, for s inside γ_1

$$\begin{aligned} (\partial_A \partial_B F)(s) &= \frac{-1}{4\pi^2} \int_{\gamma_1} \frac{1}{z-s} (zI - A)^{-1} \\ &\quad \int_{\gamma_2} \frac{1}{w-z} (wI - B)^{-1} F(w) dw dz \\ &= \frac{-1}{4\pi^2} \int_{\gamma_2} \left\{ \int_{\gamma_1} \frac{1}{w-z} \cdot \frac{1}{z-s} (zI - A)^{-1} dz \right\} \\ &\quad \cdot (wI - B)^{-1} F(w) dw \\ &= \frac{1}{2\pi j} \int_{\gamma_2} \frac{1}{w-s} \\ &\quad \cdot (wI - A)^{-1} (wI - B)^{-1} F(w) dw \end{aligned}$$

where the last step uses the fact that w is outside of γ_1 . The required result now follows from the identity

$$(A - B)(wI - A)^{-1}(wI - B)^{-1} = (wI - A)^{-1} - (wI - B)^{-1}.$$

Let $f = f_+ + f_- \in \mathcal{L}_2(-j\infty, j\infty)$, where $f_+ \in \mathcal{H}_2$ and $f_- \in \mathcal{H}_2^\perp$. Then

$$f_+(s) = \left\langle f_+(jt), \frac{1}{2\pi(jt + s^*)} \right\rangle = \left\langle f(jt), \frac{1}{2\pi(jt + s^*)} \right\rangle$$

for $\mathcal{R}e(s) > 0$. Now consider $L \in \mathbb{C}^{n \times n}$ where $\text{spec}(L)$ is in the open right-half plane and let $g \in \mathcal{H}_2^n$. Then

$$\begin{aligned} \Pi_+(sI - L)^{-1}g(s) &= \int_{-\infty}^{\infty} \frac{1}{2\pi(jt + s^*)} (jtI - L)^{-1}g(jt) dt \\ &= \frac{-1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{1}{z-s} (zI - L)^{-1}g(z) dz \\ &= (\partial_L g)(s). \end{aligned} \quad (37)$$

Let $L = PJP^{-1}$ be a Jordan decomposition with k Jordan blocks of size n_i and corresponding eigenvalues μ_i , and accordingly let

$$P = (V_1, \dots, V_k) \quad \text{and} \quad P^{-1} = \begin{pmatrix} U_1^* \\ \vdots \\ U_k^* \end{pmatrix}.$$

Then

$$(sI - L)^{-1} = \sum_{i=1}^k V_i \left\{ \sum_{\ell=1}^{n_i} \frac{1}{(s - \mu_i)^\ell} Z_{n_i}^{\ell-1} \right\} U_i^* \quad (38)$$

where Z_m is an $m \times m$ matrix with ones on the first superdiagonal and zeros elsewhere. If γ_1 is a simple closed curve containing s and μ , and γ_R is a circle of radius R about zero, then

$$\begin{aligned} \frac{1}{2\pi j} \int_{\gamma_1} \frac{1}{(z-s)} \frac{1}{(z-\mu)^\ell} dz \\ = \lim_{R \rightarrow \infty} \frac{1}{2\pi j} \int_{\gamma_R} \frac{1}{(z-s)} \frac{1}{(z-\mu)^\ell} dz = 0 \end{aligned}$$

for $\ell \geq 1$. Thus

$$\begin{aligned} \frac{1}{2\pi j} \int_{\gamma_1} \frac{1}{(z-s)} \frac{1}{(z-\mu)^\ell} f(z) dz \\ = \frac{1}{2\pi j} \int_{\gamma_1} \frac{1}{(z-s)} \frac{1}{(z-\mu)^{\ell-1}} (\partial_\mu f)(z) dz \\ = \frac{1}{2\pi j} \int_{\gamma_1} \frac{1}{(z-s)} \frac{1}{(z-\mu)} (\partial_\mu^{\ell-1} f)(z) dz \\ = (\partial_\mu^\ell f)(s). \end{aligned}$$

Hence, from (38)

$$(\partial_L F)(s) = \sum_{i=1}^k \sum_{\ell=1}^{n_i} V_i Z_{n_i}^{\ell-1} U_i^* (\partial_\mu^\ell F)(s). \quad (39)$$

VII. GENERAL CONTROLLER FORMULA

In this section we will generalize the formula of Proposition 3 to the case where $P(s) = e^{-s\tau} P_0(s)$, and $P_0(s) = C(sI - A_0)^{-1}B$ is (a minimal realization) of degree n . We recall the notation of Section III, namely $A = A_0 + EC$, $E = -R C^*$, with R, S, H defined as in (5), (15), (11).

To carry out the derivations of this section, a mild genericity assumption will be needed on $P_0(s)$, namely that the reflections of the zeros of $P_0(s)$ about the imaginary axis are disjoint from the poles of $P_0(s)$. As shown in the lemma below, this is equivalent to the spectrum of H and A being disjoint, which is the assumption we will need in the computations below. A discussion on the validity of the formula when the assumption fails is given in Remark 2.

Lemma 1: Let A, B, C , and $P_0(s)$ be defined as above, and let H be defined as in (11). Let $\{\zeta_i : i = 1, \dots, k\}$ be the zeros of $P_0(s)$ and let $\{\pi_i : i = 1, \dots, \ell\}$ be the poles of $P_0(s)$ (since $P_0(s)$ is a scalar function these two sets are disjoint). Then the following statements are equivalent:

- a) $\{\zeta_i : i = 1, \dots, k\} \cap \{-\bar{\pi}_i : i = 1, \dots, \ell\} = \emptyset$.
- b) (A, B) and (A, E) are controllable.
- c) $\text{spec}(H) \cap \text{spec}(A) = \emptyset$.

Proof: Define $\{\zeta_i^+ : i = 1, \dots, k^+\}$ and $\{\zeta_i^- : i = 1, \dots, k^-\}$ to be the set of zeros of $P_0(s)$ in $\mathcal{R}e(s) > 0$ and $\mathcal{R}e(s) < 0$ respectively. Similarly, $\{\pi_i^+ : i = 1, \dots, \ell^+\}$ and $\{\pi_i^- : i = 1, \dots, \ell^-\}$ define the set of poles of $P_0(s)$ in the same regions. We now introduce the following subsidiary statements:

- a₁) $\{\zeta_i^- : i = 1, \dots, k^-\} \cap \{-\bar{\pi}_i^+ : i = 1, \dots, \ell^+\} = \emptyset$.
- a₂) $\{\zeta_i^+ : i = 1, \dots, k^+\} \cap \{-\bar{\pi}_i^- : i = 1, \dots, \ell^-\} = \emptyset$.
- b₁) (A, B) is controllable.
- b₂) (A, E) is controllable.
- b₂') $C(-\bar{\alpha}I - A)^{-1}B \neq 0$ for all $\alpha \in \text{spec}(A)$.

It is clear that a) \Leftrightarrow a₁) and a₂), since imaginary axis poles or zeros of $P_0(s)$ can only belong to one of the sets in a). The lemma will be proved by establishing the following equivalences: a₁) \Leftrightarrow b₁); a₂) \Leftrightarrow b₂) \Leftrightarrow b₂') b₁), b₂') \Leftrightarrow c).

Let $P_0(s) = p(s)/q(s)$ where $p(s), q(s)$ are coprime polynomials and $q(s)$ is monic. Let $p(s)/r(s), q(s)/r(s)$ be normalized coprime factors of $P_0(s)$ over \mathcal{H}_∞ (so $r(s)$ is a monic polynomial of degree n).

$a_1) \Leftrightarrow b_1)$. Assume that (A, B) is not controllable. Then for some α ($\mathcal{R}e(\alpha) < 0$), $r(\alpha) = p(\alpha) = 0$. Since $p^*p + q^*q = r^*r$ then $q^*(\alpha)q(\alpha) = 0$. Therefore $q^*(\alpha) = \overline{q(-\bar{\alpha})} = 0$ since $p(s)$ and $q(s)$ are coprime. Thus $a_1)$ fails. Conversely, if $a_1)$ fails then for some α ($\mathcal{R}e(\alpha) < 0$), $p(\alpha) = q(-\bar{\alpha}) = 0$. This means that $r^*(\alpha)r(\alpha) = 0$ which implies $r(\alpha) = 0$, since $r(s)$ is Hurwitz. Therefore, $p(s)$ and $r(s)$ have a common factor. Since (C, A_0) is observable, so is (C, A) , which means that (A, B) is not controllable.

$a_2) \Leftrightarrow b_2)$. The proof is the same as for $a_1) \Leftrightarrow b_1)$ but with p and q interchanged and B and E interchanged.

$a_2) \Leftrightarrow b_2')$. Suppose $b_2')$ fails. It follows that there exists an α ($\mathcal{R}e(\alpha) < 0$) such that $r(\alpha) = p(-\bar{\alpha}) = 0$. Since $p^*p + q^*q = r^*r$, it follows that $q^*(\alpha)q(\alpha) = 0$ and hence, since $p(s)$ and $q(s)$ are coprime, that $q^*(\alpha) \neq 0$ and $q(\alpha) = 0$. Thus $a_2)$ fails. Conversely, if $a_2)$ fails, then we can write $P_0(s) = (s - \bar{\alpha})/(s + \alpha)P_1(s)$ where $\bar{\alpha}$ is a zero of $p(s)$, α is a zero of $q(s)$, and hence $P_1(s)$ has degree $n - 1$. We now find a normalized coprime factorization $P_1(s) = N_1(s)M_1(s)^{-1}$ over \mathcal{H}_∞ . Then $P_0(s) = ((s - \bar{\alpha})/(s + \alpha)N_1(s))M_1(s)^{-1}$ is a normalized coprime factorization. Therefore, since the degree of $((s - \bar{\alpha})/(s + \alpha)N_1(s), M_1(s))$ must be n , $-\alpha$ is in the spectrum of the matrix A . Finally, since normalized coprime factorizations are unique modulo a unitary constant then $b_2')$ fails.

$b_1), b_2') \Leftrightarrow c)$. Let $\alpha \in \text{spec}(A)$. Then $\alpha I + A^*$ is nonsingular since A is Hurwitz. Thus

$$\begin{aligned} \det(\alpha I - H) &= \det \begin{pmatrix} \alpha I - A & -\lambda^{-1}BB^* \\ \lambda^{-1}C^*C & \alpha I + A^* \end{pmatrix} \\ &= \det(\alpha I - A + \lambda^{-2}BB^*(\alpha I + A^*)^{-1}C^*C) \\ &\quad \cdot \det(\alpha I + A^*). \end{aligned} \quad (40)$$

Clearly

$$\begin{aligned} \det(\alpha I - H) \neq 0 &\Rightarrow B^*(\alpha I + A^*)^{-1}C^* \neq 0 \\ &\Leftrightarrow C^*(-\bar{\alpha}I - A)^{-1}B \neq 0. \end{aligned}$$

Thus $c) \Rightarrow b_2')$. Also, if (A, B) is not controllable, then there exists an $\alpha \in \text{spec}(A)$ such that $v^T(\alpha I - A) = 0$, $v^T B = 0$ for some $v \neq 0$. This implies by (40) that $\alpha \in \text{spec}(H)$. Thus $c) \Rightarrow b_1)$. Conversely, let $b_1), b_2')$ hold. Take $\alpha \in \text{spec}(A)$ and note that $\beta := \lambda^2 B^*(\alpha I + A^*)^{-1}C^* \neq 0$. We now claim that $\alpha I - A + \beta BC^*$ is nonsingular. To see this, note that $C^*(sI - A)^{-1}B = p(s)/q(s) = p(s)/\det(sI - A)$ (since $q(s)$ is monic) and that $p(s)$ and $\det(sI - A)$ have no common factor (since (A, B) is controllable and (C, A) is observable). Since

$$\begin{aligned} \det(1 + \beta C^*(sI - A)^{-1}B) \\ = \det(sI - A)^{-1} \det(sI - A + \beta BC^*) \end{aligned}$$

it follows that $\det(sI - A + \beta BC^*) = \det(sI - A) + \beta p(s)$, and hence that $\det(\alpha I - A + \beta BC^*) = \beta p(\alpha) \neq 0$. Therefore, $\alpha \notin \text{spec}(H)$ using (40). This proves $c)$. \square

We now follow the steps P1)–P3) outlined in Section III. Let λ be the largest solution of $\det L(\lambda) = 0$ in $[0, 1]$ and let $w_0 \neq 0$ solve $L(\lambda)w_0 = 0$. Then

$$v(t) = (I, 0)e^{Ht} \begin{pmatrix} \lambda^{-1}R \\ I \end{pmatrix} w_0$$

on $[0, \tau]$. This gives

$$\lambda x(t) = \begin{pmatrix} (C', 0)e^{Ht} & -\lambda^{-1}R \\ I & I \end{pmatrix} w_0 \quad \text{on } [0, \tau],$$

$$C'e^{A(t-\tau)}(I, 0)e^{H\tau} \begin{pmatrix} \lambda^{-1}R \\ I \end{pmatrix} w_0 \quad \text{on } [\tau, \infty)$$

by (9). Direct calculation shows that

$$\int_0^\tau e^{-st} e^{Ht} dt = -\partial_H(e^{-s\tau})e^{H\tau}$$

and

$$\int_\tau^\infty e^{-st} e^{A(t-\tau)} dt = e^{-s\tau}(sI - A)^{-1}.$$

Thus

$$\begin{aligned} \lambda \hat{x} &= C'(e^{-s\tau}(sI - A)^{-1}(I, 0) - (I, 0)\partial_H(e^{-s\tau})) \\ &\quad \cdot H\tau \begin{pmatrix} \lambda^{-1}R \\ I \end{pmatrix} w_0. \end{aligned} \quad (41)$$

We now assume that H is diagonalizable, for simplicity, and write $H = \sum_i w_i v_i^* \mu_i$. (In the proof of Theorem 1, we show that the result also holds without the assumption.) Thus, the term in parentheses in (41) can be written as

$$\sum_i ((sI - A)^{-1}e^{-s\tau}(I, 0) - \partial_{\mu_i}(e^{-s\tau})(I, 0))w_i v_i^*.$$

From (34) we see that

$$\begin{aligned} \partial_{\mu_i}((sI - A)^{-1}e^{-s\tau}) \\ = (\mu_i I - A)^{-1}(\partial_{\mu_i}(e^{-s\tau})I - (sI - A)^{-1}e^{-s\tau}) \end{aligned}$$

which gives

$$\lambda \hat{x} = -C' \left(\sum_i (\mu_i I - A) \partial_{\mu_i}((sI - A)^{-1}e^{-s\tau} X_i) \right)$$

where

$$X_i := (I, 0)w_i v_i^* e^{H\tau} \begin{pmatrix} \lambda^{-1}R \\ I \end{pmatrix} w_0.$$

We will now calculate $u = H_+ F^* \lambda x$. Then

$$\begin{aligned} u_1 &= -H_+ e^{s\tau} B^*(sI + A^*)^{-1} C^* C \\ &\quad \cdot \sum_i (\mu_i I - A) \partial_{\mu_i}((sI - A)^{-1}e^{-s\tau} X_i) \\ &= -\sum_i \partial_{\mu_i}(H_+ B^*(sI + A^*)^{-1} \\ &\quad \cdot C^* C(sI - A)^{-1}(\mu_i I - A)X_i) \end{aligned}$$

by Proposition 2. (Note that, with the aid of a partial fraction expansion, the proposition can be applied elementwise.) Using (15) we have

$$(sI + A^*)^{-1} C^* C(sI - A)^{-1} = -S(sI - A)^{-1} + (sI + A^*)^{-1} S.$$

Thus

$$\begin{aligned} \hat{u}_1 &= \sum_i \partial_{\mu_i}(B^* S(sI - A)^{-1}(\mu_i I - A)X_i) \\ &= -\sum_i B^* S(sI - A)^{-1} X_i \\ &= -B^* S(sI - A)^{-1} v(\tau). \end{aligned} \quad (42)$$

Now define

$$g_i(s) := C(\mu_i I - A)(sI - A)^{-1} e^{-s\tau} X_i.$$

Then

$$\begin{aligned} -\hat{u}_2 &= H_+(1 - E^*(sI + A^*)^{-1}C^*) \sum_i \partial_{\mu_i}(g_i(s)) \\ &= \sum_i \partial_{\mu_i}(g_i(s) - E^*H_+(sI + A^*)^{-1}C^*g_i(s)) \\ &= \sum_i \partial_{\mu_i}(g_i(s) - E^*\partial_{-A^*}C^*g_i(s)) \\ &= \sum_i (\partial_{\mu_i}g_i(s) - E^*(\mu_i I + A^*)^{-1} \\ &\quad \cdot (\partial_{\mu_i} - \partial_{-A^*})(C^*g_i(s))) \\ &= \sum_i ((1 - E^*(\mu_i I + A^*)^{-1}C^*)\partial_{\mu_i}(g_i(s))) \\ &\quad + E^*\partial_{-A^*} \left(\sum_i (\mu_i I + A^*)^{-1}C^*g_i(s) \right) \end{aligned} \quad (43)$$

Proposition 4: The following identity holds

$$\sum_i (\mu_i I + A^*)^{-1}C^*g_i(s) = 0.$$

In the proof of the proposition we will make use of the following fact.

Lemma 2: Let $X, Y \in \mathbb{C}^{n \times n}$ be such that

$$\text{rank}(X) + \text{rank}(Y) = n$$

and $X + Y$ be nonsingular. Then

$$Y(X + Y)^{-1}X = 0.$$

Proof: We assume that $0 < r := \text{rank}(X) < n$, otherwise the statement is trivial. We now write $X = X_1X_2$ and $Y = Y_1Y_2$, where $X_1 \in \mathbb{C}^{n \times r}$, $X_2 \in \mathbb{C}^{r \times n}$, $Y_1 \in \mathbb{C}^{n \times (n-r)}$ and $Y_2 \in \mathbb{C}^{(n-r) \times n}$. Thus

$$\begin{aligned} Y(X + Y)^{-1}X &= Y_1Y_2 \left((X_1, Y_1) \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix} \right)^{-1} X_1X_2 \\ &= Y_1Y_2 \begin{pmatrix} X_2 \\ Y_2 \end{pmatrix}^{-1} (X_1, Y_1)^{-1} X_1X_2 \\ &= Y_1(0_{(n-r) \times r}, I_{(n-r) \times (n-r)}) \begin{pmatrix} I_{r \times r} \\ 0_{(n-r) \times r} \end{pmatrix} X_2 \\ &= 0. \end{aligned} \quad \square$$

Proof of Proposition 4 It is sufficient to show that

$$\sum_i U_i T V_i = 0 \quad (44)$$

for an arbitrary constant matrix T , where

$$\begin{aligned} U_i &= (\mu_i I + A^*)^{-1}, \\ V_i &= (\mu_i I - A)(I, 0)w_i v_i^* e^{H\tau} \begin{pmatrix} \lambda^{-1}R \\ I \end{pmatrix} w_0. \end{aligned}$$

In fact it is sufficient to prove (44) for an arbitrary T of rank one, since the general fact then follows by linear superposition.

We first take $T = z_j y^*$, where z_j is an eigenvector of A^* , i.e., $A^* z_j = \bar{\gamma}_j z_j$, and y^* is an arbitrary row vector. Observe that

$$e^{H\tau} \begin{pmatrix} \lambda^{-1}R \\ I \end{pmatrix} w_0 = \begin{pmatrix} v(\tau) \\ w(\tau) \end{pmatrix} = \begin{pmatrix} I \\ \lambda^{-1}S \end{pmatrix} v(\tau) \quad (45)$$

from (12)–(14). Thus

$$\begin{aligned} \sum_i U_i T V_i &= z_j y^* \left(\sum_i \frac{1}{\mu_i + \bar{\gamma}_j} (\mu_i I - A)(I, 0)w_i v_i^* \right) \\ &\quad \cdot \begin{pmatrix} I \\ \lambda^{-1}S \end{pmatrix} v(\tau) \\ &= z_j y^* \left((I, 0) \sum_i \frac{\mu_i}{\mu_i + \bar{\gamma}_j} w_i v_i^* - (A, 0) \right. \\ &\quad \cdot \left. \sum_i \frac{1}{\mu_i + \bar{\gamma}_j} w_i v_i^* \right) \begin{pmatrix} I \\ \lambda^{-1}S \end{pmatrix} v(\tau) \\ &= z_j y^* ((I, 0)H(H + \bar{\gamma}_j I)^{-1} \\ &\quad - (A, 0)(H + \bar{\gamma}_j I)^{-1}) \begin{pmatrix} I \\ \lambda^{-1}S \end{pmatrix} v(\tau) \\ &= z_j y^* ((0, \lambda^{-1}BB^*)(H + \bar{\gamma}_j I)^{-1} \\ &\quad \cdot \begin{pmatrix} I \\ \lambda^{-1}S \end{pmatrix} v(\tau)) \end{aligned} \quad (46)$$

Since, by assumption, $-\bar{\gamma}_j$ is distinct from the eigenvalues of H , we have

$$\begin{aligned} (H + \bar{\gamma}_j I)^{-1} &= \begin{pmatrix} A + \bar{\gamma}_j I & \lambda^{-1}BB^* \\ -\lambda^{-1}C^*C & -A^* + \bar{\gamma}_j I \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \Delta^{-1} \lambda^{-1}C^*C^* (A + \bar{\gamma}_j I)^{-1} & \Delta^{-1} \end{pmatrix} \end{aligned}$$

where $\Delta = (-A^* + \bar{\gamma}_j I) + \lambda^{-2}C^*C(A + \bar{\gamma}_j I)^{-1}BB^*$. Note that Δ is nonsingular since $\det(H + \bar{\gamma}_j I) \neq 0$. Thus

$$\begin{aligned} (0, \lambda^{-1}BB^*)(H + \bar{\gamma}_j I)^{-1} \begin{pmatrix} I \\ \lambda^{-1}S \end{pmatrix} &= \lambda^{-2}BB^*\Delta^{-1}(C^*C + S(A + \bar{\gamma}_j I))(A + \bar{\gamma}_j I)^{-1} \\ &= \lambda^{-2}BB^*\Delta^{-1}(-A^* + \bar{\gamma}_j I)S(A + \bar{\gamma}_j I)^{-1}. \end{aligned}$$

We now claim that $B^*\Delta^{-1}(-A^* + \bar{\gamma}_j I) = 0$. It suffices to show that

$$\lambda^{-2}C^*C(A + \bar{\gamma}_j I)^{-1}BB^*\Delta^{-1}(-A^* + \bar{\gamma}_j I) = 0 \quad (47)$$

since $C(A + \bar{\gamma}_j I)^{-1}B \neq 0$ by Lemma 1 and C^* is a nonzero column vector. To show (47) let $Y = \lambda^{-2}C^*C(A + \bar{\gamma}_j I)^{-1}BB^*$ and $X = (-A^* + \bar{\gamma}_j I)$. Observe that $\Delta = X + Y$ is nonsingular and $\text{rank}(X) + \text{rank}(Y) = \text{rank}(\Delta)$. This follows since X is rank deficient and Y has rank at most one (B, C are vectors). Thus the conditions of Lemma 2 are satisfied and so $Y(X + Y)^{-1}X = 0$. This shows that (47) and hence (46) is zero. Since y^* is arbitrary in (46) and z_j is any eigenvector of A^* this proves that (44) holds for any $n \times n$ matrix T having range the span of the eigenvectors of A^* . This completes the proof if A is diagonalizable.

The same conclusion holds in the nondiagonalizable case as well. To see this consider a sequence of matrices A_k ($k = 1, 2, \dots$) tending to A , each of which is diagonalizable, and

let $P_k(s)$ ($k = 1, 2, \dots$) denote the corresponding perturbed transfer functions. It can be easily seen that $P_k(s) \rightarrow P(s)$ in the gap metric (this follows because the graph symbol F depends continuously on A ; see [16] for the definition and expressions for the gap metric). Further, due to the metric property of the gap (see, e.g. [16]) the optimal robustness radius $b_{\text{opt}}(P)$ depends continuously on $P(s)$. Since the largest root of $\det(L(\lambda)) = 0$ is $\lambda = \sqrt{1 - b_{\text{opt}}^2(P)}$, λ depends continuously on A as well. Clearly R and S depend continuously on A , hence so does $L(\lambda)$. Note that $L(\lambda)$ remains singular as A varies. Now suppose that $w_{0,k}$ are unit vectors satisfying $L_k(\lambda_k)w_{0,k} = 0$, $k = 1, 2, \dots$ (where L_k is defined via (16) for the case of each $P_k(s)$, respectively). Without loss of generality we can take $w_{0,k}$ to be a convergent sequence (otherwise we select an appropriate subsequence), and let the limit point be w_0 . Then $L(\lambda)w_0 = 0$, so this limit point can be taken as the definition of w_0 . Since H is assumed diagonalizable and since it depends continuously on A , we can assume that, in each case, H_k is also diagonalizable (possibly for k sufficiently large). Further, we can select eigenvectors for H_k tending to corresponding ones for H . Compiling the above, it is seen that Σ, U, TV_z depends continuously on the perturbations of A . By the earlier part of the proof this expression is identically zero in the case of each A_k . Thus in the limit it is also zero for the given $P(s)$ and A . This completes the proof. \square

Proposition 4 and (43) show that

$$u_2 = - \sum_i (1 - E^*(\mu_i I + A^*)^{-1} C^*) \partial_{\mu_i}(g_i(s)) \quad (48)$$

$$= - \sum_i M^*(\mu_i) \partial_{\mu_i}(g_i(s)). \quad (49)$$

We now expand $\partial_{\mu_i}(g_i(s))$ using (35) so that the ∂ -operator acts only on the infinite-dimensional element $e^{-s\tau}$

$$\begin{aligned} \partial_{\mu_i}(g_i(s)) &= C(\mu_i I - A) \partial_{\mu_i}((sI - A)^{-1} e^{-s\tau} X_i) \\ &= C(\mu_i I - A)(sI - A)^{-1} (\partial_{\mu_i}(e^{-s\tau}) I \\ &\quad - (\mu_i I - A)^{-1} e^{-\mu_i \tau} X_i) \\ &= C(sI - A)^{-1} (\mu_i I - A) X_i \partial_{\mu_i}(e^{-s\tau}) \\ &\quad - C(sI - A)^{-1} X_i e^{-\mu_i \tau}. \end{aligned}$$

Using the following equalities

$$\begin{aligned} &\sum \mu_i M^*(\mu_i) \partial_{\mu_i}(e^{-s\tau}) X_i \\ &= (I, 0) H M^*(H) \partial_H(e^{-s\tau}) e^{H\tau} \begin{pmatrix} \lambda^{-1} R \\ I \end{pmatrix} w_0, \\ &\sum A M^*(\mu_i) \partial_{\mu_i}(e^{-s\tau}) X_i \\ &= (A, 0) M^*(H) \partial_H(e^{-s\tau}) e^{H\tau} \begin{pmatrix} \lambda^{-1} R \\ I \end{pmatrix} w_0, \\ &\sum M^*(\mu_i) X_i e^{-\mu_i \tau} \\ &= (I, 0) M^*(H) \begin{pmatrix} \lambda^{-1} R \\ I \end{pmatrix} w_0 \end{aligned}$$

and (49) we obtain

$$\begin{aligned} \hat{u}_2 &= C(sI - A)^{-1} (I, 0) M^*(H) \begin{pmatrix} \lambda^{-1} R \\ I \end{pmatrix} w_0 \\ &\quad - C(sI - A)^{-1} B(0, \lambda^{-1} B^*) M^*(H) \partial_H(e^{-s\tau}) \\ &\quad \cdot \begin{pmatrix} I \\ \lambda^{-1} S \end{pmatrix} v(\tau). \end{aligned} \quad (50)$$

We are now ready to state the form of the optimal compensator.

Theorem 1: Let $P(s) = e^{-s\tau} P_0(s)$ be a single-input/single-output plant with $P_0(s) = C(sI - A_0)^{-1} B$ a minimal realization and $\tau > 0$. Suppose that $P_0(s)$ satisfies condition a) of Lemma 1. Then the optimal robustness radius for gap ball uncertainty is given by the formula

$$b_{\text{opt}}(P) = (1 - \lambda_{\text{max}}^2)^{1/2}$$

where λ_{max} is the largest solution of the transcendental equation $\det L(\lambda) = 0$ on $[0, 1]$ (which is defined by (16)). The transfer function of the optimal controller is given by

$$K_{\text{opt}}(s) = \frac{B^* S(sI - A)^{-1} B_1}{C(sI - A)^{-1} B_2 - C(sI - A)^{-1} B h(\tau, s)} \quad (51)$$

where $L(\lambda)w_0 = 0$

$$w_1 = \begin{pmatrix} \lambda^{-1} R \\ I \end{pmatrix} w_0,$$

$$B_1 = (I, 0) e^{H\tau} w_1,$$

$$B_2 = (I, 0) M^*(H) w_1,$$

$$h(\tau, s) = (0, \lambda^{-1} B^*) M^*(H) \partial_H(e^{-s\tau}) e^{H\tau} w_1$$

with R, S, H defined as in (5), (15), (11), and $\lambda = \lambda_{\text{max}}$ throughout.

Before proving the theorem we will establish the following result.

Lemma 3: Consider any (nonzero) strictly proper rational function $P_0(s)$. Then $b_{\text{opt}}(e^{-s\tau} P_0(s))$ is a strictly decreasing, continuous function of τ .

Proof: Let $P_0 = N_0/M_0$ be a normalized coprime factorization over \mathcal{H}_∞ . Then $(e^{-s\tau} N, M)$ varies continuously with τ in the \mathcal{H}_∞ norm. Thus $\| \Pi_- \begin{pmatrix} e^{-s\tau} N_0^* \\ M_0^* \end{pmatrix} \|_{\mathcal{H}_2}$ varies continuously with τ , and the same is true for $b_{\text{opt}}(e^{-s\tau} P_0(s))$ using (2). Thus, the statement of the lemma requires proof of the following fact

$$\Pi_- \begin{pmatrix} -e^{-s\tau_2} N_0^* \\ M_0^* \end{pmatrix}_{\mathcal{H}_2} > \Pi_- \begin{pmatrix} -e^{-s\tau_1} N_0^* \\ M_0^* \end{pmatrix}_{\mathcal{H}_2} \quad (52)$$

for $\tau_2 > \tau_1$ (cf. (2)). We will prove a slightly more general fact than (52). Let $F \in \mathcal{H}_\infty^{1 \times 2}$ be continuous (and non-constant) on the compactified right-half plane, co-inner and right invertible over \mathcal{H}_∞ . Let $\Phi(s) \in \mathcal{H}_\infty^{2 \times 2}$ be unitary ($\Phi^* \Phi = \Phi \Phi^* = I$) with $\det \Phi(s)$ not a constant. We will show that $\| \Pi_- \Phi^* F^* \|_{\mathcal{H}_2} > \| \Pi_- F^* \|_{\mathcal{H}_2}$. First note that, for any $\hat{x} \in \mathcal{H}_2$

$$\begin{aligned} \| \Pi_- \Phi^* F^* \hat{x} \|_2^2 &= \| \Pi_- \Phi^* \Pi_- F^* \hat{x} + \Pi_- \Phi^* \Pi_+ F^* \hat{x} \|_2^2 \\ &= \| \Pi_- \Phi^* \Pi_- F^* \hat{x} \|_2^2 + \| \Pi_- \Phi^* \Pi_+ F^* \hat{x} \|_2^2 \\ &= \| \Pi_- F^* \hat{x} \|_2^2 + \| \Pi_- \Phi^* \Pi_+ F^* \hat{x} \|_2^2. \end{aligned} \quad (53)$$

The second equality follows since the inner product

$$\langle \Pi_- \Phi^* \Pi_- F^* \hat{x}, \Pi_- \Phi^* \Pi_+ F^* \hat{x} \rangle = 0.$$

We now take \hat{x} to be the maximal vector for $\Pi_- F^*|_{\mathcal{H}_2}$ and write $\hat{v} = \Pi_+ F^* \hat{x}$. Let us suppose that $\Pi_- \Phi^* \hat{v} = 0$, i.e., the second term in (53) vanishes. This means that $\hat{v} = \Phi \hat{w}$ for some $\hat{w} \in \mathcal{H}_2$, so \hat{v} has a nontrivial inner factor. But this contradicts the fact that \hat{x} is outer and \hat{v}/\hat{x} is left invertible over \mathcal{H}_∞ for the assumed conditions on F [19]. This establishes the required result. \square

Proof of Theorem 1: Under the assumption that $H = H_\lambda$ is diagonalizable the result follows from Proposition 1, (2), and steps P1)–P3). It follows that $K_{\text{opt}}(s) = -\hat{u}_1/\hat{u}_2$, where \hat{u}_1 and \hat{u}_2 are given in (42) and (50). To show that the formula is also valid when H is not diagonalizable we use a perturbation argument. We break the reasoning into two steps.

Step 1: We claim that H_λ has distinct eigenvalues for almost all λ if the conditions of Lemma 1 hold. To see this we write $C(sI - A)^{-1}B = p(s)/r(s)$ where $p(s)$ and $r(s)$ are coprime and $r(s)$ has degree n (this is possible since (A, B) is controllable). We note that p^*p and r^*r are coprime (this follows as in the proof of $a_1 \leftrightarrow b_1$) in Lemma 1). Thus $\lambda^2 r^*r - p^*p$ has distinct roots for almost all λ (this follows by taking derivatives with respect to s , eliminating λ^2 to give a nonzero polynomial in s , and verifying that these root locations can only be achieved by finitely many λ). The claim now follows by noting that the eigenvalues of H_λ coincide with the zeros of the expression

$$1 - \lambda^{-2} C(sI - A)^{-1} B B^* (-sI - A^*)^{-1} C^*$$

which are the same as the zeros of $\lambda^2 r^*r - p^*p$.

Step 2: Suppose for some τ and $\lambda = \lambda_{\max}(\tau)$ that H_λ has repeated roots. Take a sequence τ_i ($i = 1, 2, \dots$) with $\tau_i < \tau$ and $\tau_i \rightarrow \tau$. By Lemma 3, $\lambda_i = \lambda_{\max}(\tau_i) < \lambda$ and $\lambda_i \rightarrow \lambda$. Using Step 1 we can assume, without loss of generality, that H_{λ_i} has distinct eigenvalues. Now suppose that w_i are unit vectors satisfying $L(\lambda_i)w_i$. Without loss of generality we can take w_i to be a convergent sequence (otherwise we select an appropriate subsequence), and let the limit point be w_0 . Then $L(\lambda)w_0 = 0$. Let \hat{x} be defined as in (41) with $H = H_\lambda$, and let \hat{x}_i be defined by the same equation but with w_0 replaced by w_i , τ replaced by τ_i and H by H_{λ_i} . We note that $\hat{x}_i \rightarrow \hat{x}$. Define $\hat{u} = \Pi_+ F^* \lambda \hat{x}$ and let $(\hat{u})_i$ be defined similarly but with \hat{x} replaced by \hat{x}_i and τ replaced by τ_i in F . Then $(\hat{u})_i \rightarrow \hat{u}$. The proof is now completed by noting that, with the first and second components of \hat{u} defined through (42) and (50), we also have $(\hat{u})_i \rightarrow \hat{u}$ by virtue of the fact that $M^*(H)$ and $\partial_H(e^{-s\tau})$ are continuous functions of H . Thus (42) and (50) are correct expressions even when H is not diagonalizable. \square

Remark 1: The form of $K_{\text{opt}}(s)$ given in (51) is the ratio of two \mathcal{H}_2 functions. Consequently, it is not a coprime factorization over \mathcal{H}_∞ since there will be a common zero at infinity. In fact it was proved in [19] that $(s+1)\hat{x}$ is a unit in \mathcal{H}_∞ and that \hat{u}/\hat{x} belongs to \mathcal{H}_∞ and is left invertible. Thus (51) becomes a coprime factorization over \mathcal{H}_∞ on multiplication of the numerator and denominator by $(s+1)$.

Remark 2: (On the failure of conditions of Lemma 1.) If the poles and zeros of $P_0(s)$ do not satisfy the genericity assumption a) of Lemma 1 then $M^*(H)$ may not be defined. Examples where such a situation arises are: $P_0(s) = (s-1)/(s(s+1))$ and $P_0(s) = (s+1)/(s(s-1))$. It seems to be the case in such situations, however, that $(I, 0)M^*(H)w_1$, as well as the second term in (50), is bounded in the limit as $P_0(s)$ is approached. If a well-defined limit can be proved to exist then the controller formula of Theorem 1 would also be valid in this sense when Lemma 1 fails, by virtue of known results about the continuity of \hat{u}/\hat{x} (see [25]).

Remark 3: It should be observed that the infinite dimensionality of $K_{\text{opt}}(s)$ is confined to the term $h(\tau, s)$, and more specifically to $\partial_H(e^{-s\tau})$. It is possible to approximate $\partial_H(e^{-s\tau})$ uniformly in the \mathcal{H}_∞ norm, which then implies that the corresponding finite dimensional controller converges to $K_{\text{opt}}(s)$ in the gap metric. (This follows since the corresponding coprime factors then converge in the \mathcal{H}_∞ norm, cf. Remark 1.) To approximate $\partial_H(e^{-s\tau})$ uniformly, it suffices to find a sequence of functions $f_k(s)$, with uniformly bounded \mathcal{H}_∞ norm, such that $e^{-s\tau}$ is approximated uniformly on any finite interval of the imaginary axis. If H has an eigenvalue of multiplicity r on the imaginary axis, however, then it is also necessary to have the first r derivatives of $f_k(s)$ tending to those of $e^{-s\tau}$ at that point. Possible choices satisfying these conditions are: $f_k(s) = 1/(1 + s\tau/k)^k$, $f_k(s) = (1 - 0.5s\tau/k)^k/(1 + 0.5s\tau/k)^k$ or general Padé approximants. An alternative approach to finite dimensional controller design for the \mathcal{H}_∞ design problem of this paper has been given in [19] which explores approximation of the time-domain vector $\iota(t)$.

Remark 4: An alternative approach to the design technique presented in this paper is to first approximate the plant $e^{-s\tau}P_0(s)$, or simply the delay element $e^{-s\tau}$, by a rational transfer function, and then design an optimal controller using finite-dimensional techniques. It is well known and an easy consequence of the metric property of the gap that the robustness radius $b_{\text{opt}}(\cdot)$ is a continuous function of the data (i.e., the system transfer function). Therefore, any controller designed using a finite-dimensional approximation of the plant model, which is close to the plant in the gap metric, will qualify as a suboptimal controller for the nominal model. The performance degradation will be reflected in a reduced value for $b_{P,K}$ and, accordingly, in an increased value for the \mathcal{H}_∞ -norm of $(\begin{smallmatrix} I \\ J \end{smallmatrix})(I + KP)^{-1}(I, K)$. In general, however, controllers obtained in this way may not be close in the gap to the optimal one. One such case where this occurs is when the maximal singular value of the operator Γ has multiplicity greater than one. Such an example is $P(s) = s/(s^2 + 1)$ (given in [17] and discussed in [32]). At the moment, sufficient conditions which guarantee the convergence of suboptimal controllers to the optimal one are not known.

VIII. DESIGN EXAMPLE

In this section we will use the formula in Theorem 1 to design a controller to achieve disturbance attenuation and robust set-point following for a delay system. We will take a plant with two lightly damped poles and a nontrivial delay.

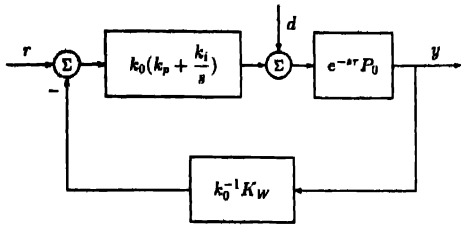


Fig. 2. Scheme for optimal robustness controller.

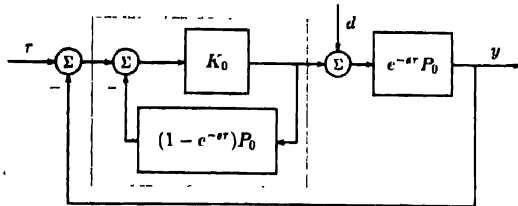
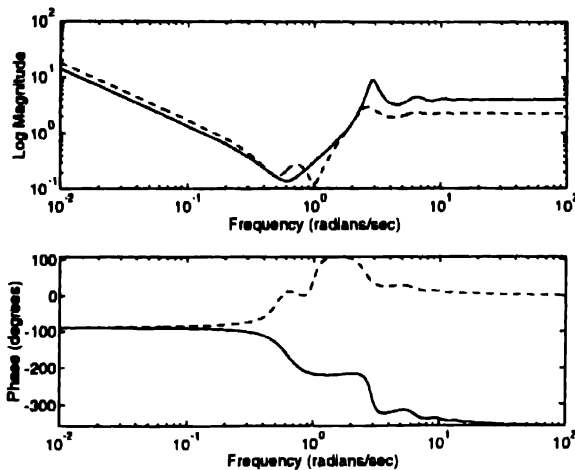


Fig. 3. Smith predictor control system.


 Fig. 4. Bode plots of optimal robustness and Smith predictor compensators: K and K_s .

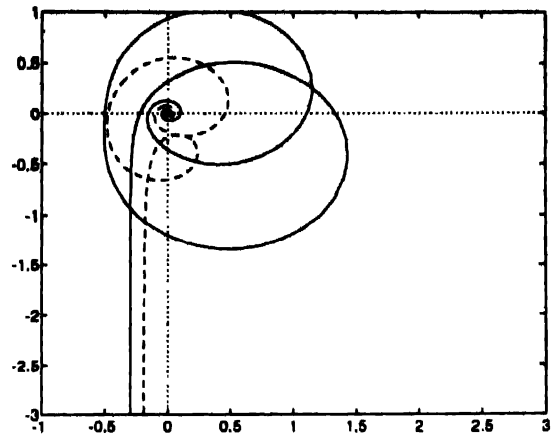
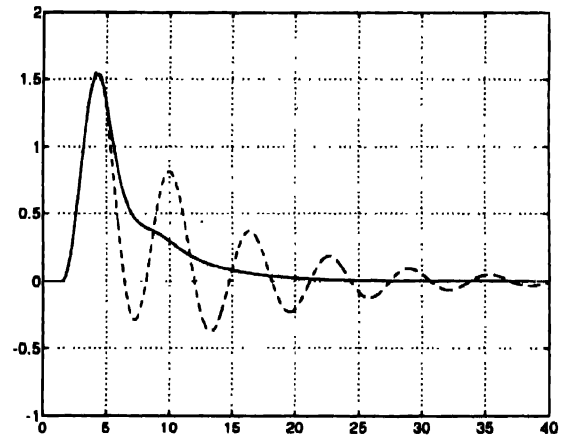
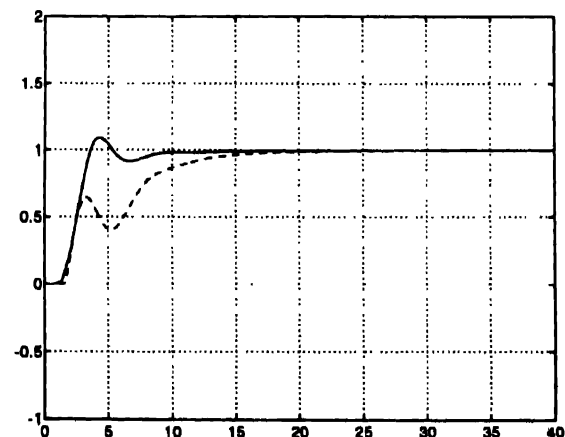
This type of system is a suitable nominal model in a number of process control applications. We will show that our design has significantly better performance than a Smith predictor.

The plant we will consider is

$$P(s) = e^{-1.5s} \frac{1}{s^2 + 0.2s + 1}.$$

We will demand increased damping of the oscillations in response to disturbances at the plant input, together with robust asymptotic tracking of constant reference inputs. To achieve these goals, we follow the H_∞ loop shaping (optimal robustness) design procedure of Glover-McFarlane [21]. We select an appropriate weighting function $W(s)$ and compute the optimal controller $K_W(s)$ for $W(s)P(s)$. The final controller for $P(s)$ is then given by $K(s) = K_W(s)W(s)$.

Since an integral term is required in the controller for robust tracking, a proportional-plus-integral weighting function $W(s) = k_p + k_i/s$ was selected. The values $k_p = 1.4$ and $k_i = 0.4$ were found to give an acceptable loop shape and a sufficiently large robustness margin of $b_{opt} = 0.3296$. (Increasing k_p and k_i gives larger loop gain but causes b_{opt}


 Fig. 5. Nyquist diagrams of return ratio transfer functions $P(s)K(s)$ and $P(s)K_s(s)$.

 Fig. 6. Response of y to a unit step at d with optimal robustness and Smith predictor compensators.

 Fig. 7. Response of y to a unit step at r with optimal robustness and Smith predictor compensators.

to decrease.) The final $K(s)$ obtained is stable (except for the pole at $s = 0$) and is nonminimum phase as evidenced by the lagging phase characteristic between 0.3 and 1.0 rad/sec (see Fig. 4). A suitable implementation of the controller is shown

in Fig. 2, where $k_0 = K_W(0)$. (Note that the specifications are also met with $K(s)$ in the forward path and unity negative feedback. The arrangement of Fig. 2, however, ensures that the response from r to y is not adversely affected by the nonminimum phaseness of $K_W(s)$. Moreover, any admissible command response—a stable system with at least second order roll-off in series with a time delay of τ seconds—can be achieved in this setup using an additional stable filter outside the loop. In particular, the command response could be made the same as the Smith predictor below.)

The idea of the Smith predictor [27] is that a precompensator K_0 is designed for P_0 to give a desired command response from r to y in the delay-free situation. The precompensator

$$K_s = \frac{K_0}{1 + (1 - e^{-s\tau})P_0K_0}$$

stabilizes the plant $e^{-s\tau}P_0$, if P_0 is stable and gives the same command response but with a delay of τ seconds (see Fig. 3). For this example, K_0 was designed using the H_∞ loop shaping (optimal robustness) procedure using the same weighting function as above: $W(s) = 1.4 + 0.4/s$. This gives

$$K_0(s) = \frac{(2.302s^2 + 0.470s + 0.612)(s + 0.286)}{s(s^2 + 2.693s + 0.719)}.$$

The Bode plots for both compensators K and K_s are shown for comparison in Fig. 4. The Nyquist plots of the return ratio transfer functions PK and PK_s are shown in Fig. 5. Although there are some broad similarities in the forms of these plots, it should be pointed out that a basic property of the Smith predictor is that open loop poles of P_0 are cancelled by the zeros of the compensator K_s . Thus, the open-loop lightly damped poles in P are not shifted by feedback in Fig. 3, in contrast to the situation in Fig. 2. The adverse effects of this cancellation become evident if the response at the plant output y is compared to step disturbances d at the plant input (see Fig. 6). We remark that the bad properties of the Smith predictor in case the plant has poles near the imaginary axis, have already been noted in the literature, and this has led to modifications of the scheme being proposed [33].

Finally, a comparison of the command response to step inputs at r is shown in Fig. 7. Note that in Figs. 4–7 the solid line denotes the optimal robustness compensator and the broken line the Smith predictor.

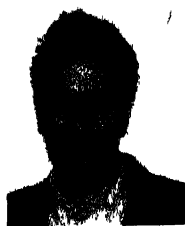
REFERENCES

- [1] H. Bercovici, C. Foias, and A. Tannenbaum, "On skew Toeplitz operators," *Operator Theory Advances and Applications*, vol. 32, pp. 21–43, 1988.
- [2] L. de Branges, "Some Hilbert spaces of analytic functions I," *Trans Amer Math Soc*, vol. 106, pp. 445–468, 1963.
- [3] L. de Branges and J. Rovnyak, "Canonical models in quantum scattering theory," in *Perturbation Theory and its Applications in Quantum Mechanics*, C. H. Wilcox, Ed. New York: Wiley, 1966.
- [4] S. Buddie, T. T. Georgiou, U. Ozguner, and M. C. Smith, "Flexible structure experiments at JPL and WPAFB," *Int J Contr*, vol. 58, no. 1, pp. 1–19, 1993.
- [5] R. F. Curtain, " \mathcal{H}_∞ control for distributed parameter systems: A survey," in *Proc 29th IEEE Conf Decis Contr*, Honolulu, HI, Dec 1990, pp. 22–26.
- [6] ———, "A synthesis of time and frequency domain methods for the control of infinite-dimensional systems: A system theoretic approach," in *Frontiers in Applied Mathematics*, H. T. Banks, Ed. Philadelphia, PA: SIAM, 1992, pp. 171–224.
- [7] H. Dym, "J-contractive matrix functions, reproducing kernel Hilbert spaces and interpolation," *Regional Conf Series Math/Conf Board Math Sci*, vol. 71, 1989.
- [8] ———, "Shifts, realizations and interpolation, redux," *Oper Theory Adv Appl*, vol. 73, pp. 182–243, 1994.
- [9] ———, "On the zeros of some continuous analogues of matrix orthogonal polynomials and a related extension problem with negative squares," *Comm Pure Appl Math*, vol. 47, pp. 207–256, 1994.
- [10] D. S. Flamm, "Control of delay systems for minimax sensitivity," Massachusetts Institute of Technology, Rep LIDS-TH-1560, 1986.
- [11] ———, "Outer factor 'absorption' for \mathcal{H}_∞ control problems," *Int J Robust Nonlinear Contr*, to appear.
- [12] D. S. Flamm and H. Yang, " \mathcal{H}_∞ optimal mixed sensitivity for general distributed systems," in *Proc 29th IEEE Conf Decis Contr*, Honolulu, HI, Dec 1990, pp. 134–139.
- [13] C. Foias, A. Tannenbaum, and G. Zames, "Weighted sensitivity minimization for delay systems," *IEEE Trans Automat Contr*, vol. AC-31, no. 8, pp. 763–766, 1986.
- [14] ———, "Some explicit formulae for the singular values of a certain Hankel operators with factorizable symbol," *SIAM J Math Analysis*, vol. 19, pp. 1081–1091, 1988.
- [15] T. T. Georgiou and M. C. Smith, "Robust control of feedback systems with combined plant and controller uncertainty," in *Proc 1990 Amer Contr Conf*, May 1990, pp. 2009–2013.
- [16] ———, "Optimal robustness in the gap metric," *IEEE Trans Automat Contr*, vol. 35, pp. 673–686, 1990.
- [17] ———, personal notes, 1990.
- [18] ———, "Robust stabilization in the gap metric: Controller design for distributed plants," *IEEE Trans Automat Contr*, vol. 37, pp. 1133–1143, 1992.
- [19] ———, "Topological approaches to robustness," in *Lecture Notes in Control and Information Sciences*, Berlin: Springer Verlag, vol. 185, 1993, pp. 222–241.
- [20] D. G. Meyer and G. F. Franklin, "A connection between normalized coprime factorizations and linear quadratic regulator theory," *IEEE Trans Automat Contr*, vol. AC-32, pp. 227–228, 1987.
- [21] D. C. McFarlane and K. Glover, "Robust controller design using normalized coprime factor plant descriptions," *Lecture Notes in Control and Information Sciences*, vol. 138, Berlin: Springer Verlag, 1989.
- [22] H. Ozbay and A. Tannenbaum, "A skew Toeplitz approach to the \mathcal{H}_∞ optimal control of multivariable distributed systems," *SIAM J Contr Optim*, vol. 28, pp. 653–670, 1990.
- [23] H. Ozbay, M. C. Smith, and A. Tannenbaum, "Mixed sensitivity optimization for a class of unstable infinite dimensional systems," *Linear Algebra and its Applications*, vol. 178, pp. 43–83, 1993.
- [24] J. R. Partington and K. Glover, "Robust stabilization of delay systems by approximation of coprime factors," *Syst Contr Lett*, vol. 14, pp. 325–331, 1990.
- [25] V. V. Peller, "Continuity properties of the operator of best approximation by analytic functions," *LOMI Preprints*, E-13–87, Leningrad, 1987.
- [26] M. Schechter, *Principles of Functional Analysis*. New York: Academic, 1971.
- [27] O. J. M. Smith, "Closer control of loops with dead time," *Chem Eng Progress*, vol. 53, pp. 217–219, 1957.
- [28] G. Tadmor, "An interpolation problem associated with \mathcal{H}_∞ -optimal design in systems with distributed input lags," *Syst Contr Lett*, vol. 8, pp. 313–319, 1987.
- [29] ———, " \mathcal{H}_∞ interpolation in systems with commensurate input lags," *SIAM J Contr Optim*, vol. 27, no. 3, pp. 511–526, 1989.
- [30] O. Tokor and H. Ozbay, "On suboptimal robustness in the gap metric for MIMO delay systems," in *Proc Amer Contr Conf*, June 1994, pp. 3183–3187.
- [31] M. Vidyasagar, "Normalized coprime factorizations for non strictly proper systems," *IEEE Trans Automat Contr*, vol. 33, pp. 300–301, 1988.
- [32] L. Y. Wang and G.-M. Joh, "Continuity of optimal robustness and robust stabilization in slowly varying systems," *Automatica*, vol. 31, pp. 1–11, 1995.
- [33] K. Watanabe, Y. Ishiyama, and M. Ito, "Modified Smith predictor control for multivariable systems with delays and unmeasurable step disturbances," *Int J Contr*, vol. 37, pp. 959–973, 1983.
- [34] G. Zames and S. K. Mitter, "A note on essential spectrum and norms of mixed Hankel-Toeplitz operators," *Syst Contr Lett*, vol. 10, pp. 159–165, 1988.
- [35] K. Zhou and P. P. Khargonekar, "On the weighted sensitivity minimization problem for delay systems," *Syst Contr Lett*, vol. 8, pp. 307–312, 1987.



Harry Dym received the B.E. degree from the Cooper Union School of Engineering, New York, in 1959 and the M.Sc. degree in electrical engineering from the California Institute of Technology, Pasadena, in 1960. He received the Ph.D. degree in mathematics from Massachusetts Institute of Technology, Cambridge, in 1965.

In 1960, Dr. Dym worked in the Communications Department of the MITRE Corp. Since 1970, he has been associated with the Department of Theoretical Mathematics of the Weizmann Institute of Science, Rehovot, Israel. His current research interests include the intersection of operator theory and complex analysis and their applications to problems of interpolation and approximation.



Tryphon T. Georgiou was born in Athens, Greece, on October 18, 1956. He received the Diploma in mechanical and electrical engineering from the National Technical University of Athens, Greece, in 1979 and the Ph.D. degree from the University of Florida, Gainesville, in 1983.

He served in the faculty of Florida Atlantic University from 1983–1986 and on the faculty of Iowa State University from 1986–1989. Since 1989, he has been with the University of Minnesota where he is currently a Professor of Electrical Engineering.

His research interests include control theory, signal processing, and applied mathematics.

Dr. Georgiou is a corecipient (with M. C. Smith) of the 1992 George Axelby Best Paper Award. He was an Associate Editor for IEEE TRANSACTIONS ON AUTOMATIC CONTROL from 1991–1992 and is currently an Associate Editor for the *SIAM Journal on Control and Optimization*.



Malcolm C. Smith (M'90) was born in Bolton, England, on June 12, 1957. He received the B.A. degree in mathematics in 1978, the M.Phil. degree in control engineering and operational research in 1979, and the Ph.D. degree in control engineering in 1982, all from Cambridge University, England.

He was subsequently a Research Fellow at the Institute for Flight Systems Dynamics, DLR, Germany, a Visiting Assistant Professor and Research Fellow with the Department of Electrical Engineering at McGill University, Canada, and an Assistant Professor with the Department of Electrical Engineering at Ohio State University, Columbus. Since 1990, he has been a lecturer in engineering at the University of Cambridge, England. His current research interests include the theory and practice of control engineering.

Dr. Smith is a corecipient (with T. T. Georgiou) of the 1992 George Axelby Best Paper Award. He is currently an Associate Editor for the *SIAM Journal on Control and Optimization*.

Stochastic System Identification with Noisy Input–Output Measurements Using Polyspectra

Jitendra K. Tugnait, *Fellow, IEEE*, and Yisong Ye, *Member, IEEE*

Abstract—Two new classes of parametric, frequency domain approaches are proposed for estimation of the parameters of scalar, linear “errors-in-variables” models, i.e., linear systems where measurements of both input and output of the system are noise contaminated. One of the proposed classes of approaches consists of linear estimators where using the bispectrum or the integrated polyspectrum (bispectrum or trispectrum) of the input and the cross-bispectrum or the integrated cross-polyspectrum (respectively, of the input–output), the system transfer function is first estimated at a number of frequencies exceeding one-half the number of unknown parameters. The estimated transfer function is then used to estimate the unknown parameters using an overdetermined linear system of equations. In the second class of approaches, quadratic transfer function matching criteria are optimized by using the results of the linear estimators as initial guesses. Both classes of the parameter estimators are shown to be consistent in any measurement noise that has symmetric probability density function when the bispectral approaches are used. The proposed parameter estimators are shown to be consistent in Gaussian measurement noise when trispectral approaches are used. The input to the system need not be a linear process but must have nonvanishing bispectrum or trispectrum. Computer simulation results are presented in support of the proposed approaches. Performance comparisons with several existing approaches based upon computer simulations are also provided.

I. INTRODUCTION

PARAMETER estimation and system identification for stochastic linear systems have been a topic of active research for over three decades now [7], [19], [20], [37], [45]. It is often assumed that the measurements of the system output are noisy but the measurements of the input to the system are perfect. The problem considered in this paper is that of identification of stochastic linear systems when the input as well as the output measurements are noisy. An interesting example of system identification with noisy input may be found in [38] where the problem of (off-line) estimation of certain parameters associated with the dynamics of a submerged undersea vehicle is studied. The various control inputs and the corresponding motion variables can only be remotely sensed and, hence, are contaminated with sensor noises. In multivariate time series problems where one is interested in exploring the relationship (transfer function) between two groups of variables, it is more logical to “symmetrically” model the system by allowing all

measured variables to be noisy [29]. Such models are called errors-in-variables models in the econometrics literature [6].

In this paper we consider a specific class of systems where the input process is non-Gaussian and the measurement noise at the input as well as the output is Gaussian if the input process has symmetric probability density function (PDF). The noise processes are allowed to be non-Gaussian with symmetric PDF if the input process has asymmetric PDF. Clearly, this model may not be always appropriate but there are several situations of practical interest where such assumptions are valid. For instance, a pseudo-random binary sequence is often used to probe a control system for identification purposes [7], [19], [20]; such sequences are clearly non-Gaussian with nonvanishing trispectrum. The problem of differential time delay estimation given measurements at two spatially separated sensors can also be cast in the framework of this paper [10]–[12], [26], [28].

Past approaches to the problem of stochastic linear system identification for errors-in-variables models may be divided into two classes: those that exploit only the second-order statistics and those that use higher (higher than second) order cumulant statistics. A good survey of the work done prior to about 1980 is given in [2]. For later work, see [3]–[9], [16]–[18], [29], and [32]. Higher order statistics have been exploited in [5], [6], [12]–[14], [31], [33]–[35], [39], [40], and [44]. Söderström [1], [2] allows only white additive noise at the input, and furthermore, the input and the output noises are assumed to be mutually uncorrelated. Most of the early work in this area has been done in econometrics. When only second-order statistics are exploited, it is known that, in general, there does not exist a unique solution [3]–[6], [9]. Therefore, attention has been focused on characterization of the class of transfer functions which fit the data.

The use of higher order cumulant statistics [23], [24] can, in principle, yield consistent parameter estimates. Deistler [5] (also [14]) has shown how to estimate the transfer function of a single-input/single-output (SISO) system in the frequency domain by use of the higher order cumulant spectrum of the output and the higher order cumulant cross-spectrum of the input–output record. Cross-cumulants between input–output have been exploited for FIR (finite impulse response) filter identification in [12] and [44]. Instrumental variable type approaches have been considered in [31] and [35] where the consistency results have been proven only for i.i.d. (independent and identically distributed) inputs. In [34] a novel cost function involving the third-order cumulants of the input–output data has been proposed, and it has been shown to

Manuscript received September 24, 1993; revised June 6, 1994. Recommended by Past Associate Editor, B. Pasik-Duncan. This work was supported in part by the National Science Foundation under Grants MIP-9101457 and MIP-9312559.

The authors are with the Department of Electrical Engineering, Auburn University, Auburn, AL 36849 USA.

IEEE Log Number 9409418.

be proportional to a conventional mean-square error criterion based upon noiseless data. Consistency of the approach of [34] has been established under several restrictive conditions such as the system input is a linear process. Also [34] requires that the noise processes, if non-Gaussian with symmetric PDF, should be linear processes. In this paper we do not require any such constraints. It should be noted that unlike the second-order statistics case, one can not, in general, model a stationary random process with a given higher order cumulant spectrum as having been generated by driving a linear system with an i.i.d. sequence [27]. In [40] several linear/iterative approaches using the auto- and/or cross- third-order cumulants of the input-output processes have been presented. Conditions under which the proposed approaches will yield consistent parameter estimators have not been provided in [40].

In [28] and [39] the square root of the magnitude of the fourth cumulant of a generalized error signal is proposed as a performance criterion for parameter estimation. Both single-input single-output and multiple-input multiple-output models have been considered in [39]. Strong consistency of the proposed parameter estimator has been established for linear inputs in [39] for Gaussian noise processes. The approach of [28] and [39] results in a nonlinear estimator that requires a good initial guess for convergence; unfortunately, no method for reliable initialization was provided in [39].

In this paper two new classes of parametric frequency domain approaches are proposed for estimation of the parameters of scalar, linear errors-in-variables models. One of the proposed approaches is a linear estimator where using the bispectrum of the input and the cross-bispectrum of the input-output, the system transfer function is first estimated at a number of frequencies exceeding one-half the number of unknown parameters. The estimated transfer function is used to estimate the unknown parameters using an overdetermined linear system of equations. In the second approach a quadratic transfer function matching criterion is optimized by using the linear estimator as an initial guess. Both the parameter estimators are shown to be consistent in any measurement noise that has symmetric PDF. The input to the system need not be a linear process but must have nonvanishing bispectrum. These two classes of approaches can be modified to exploit integrated polyspectrum, either bispectrum or trispectrum. The integrated polyspectrum is defined as a cross-spectrum between the process and a nonlinear function of the process; see Section II for further details. As discussed in Section II, integrated polyspectrum (bispectrum or trispectrum) is computed as a cross-spectrum; hence, it is computationally cheaper than the corresponding polyspectrum particularly in the case of the fourth-order cumulant spectrum. If the non-Gaussian input to the linear system has a symmetric PDF, its bispectrum and the integrated bispectrum will vanish whereas its trispectrum and the integrated trispectrum will not, provided that the fourth cumulant of the input γ_{4u} is nonzero. Extension of the bispectrum-based approaches to trispectrum-based approaches is computationally complex, and the resulting estimators are likely to have poor statistical performance because of the high variance of the trispectrum estimators [24]. Herein lies the significance of the integrated polyspectrum-based approaches

which apply with almost equal computational and programming ease to both cases, those involving integrated bispectrum as well as those concerned with integrated trispectrum.

The paper is organized as follows. In Section II, a more precise statement of parameter estimation problem under consideration is provided along with a definition and some analysis of the integrated polyspectrum of interest in this paper. The bispectrum-based approaches are described in Section III. The integrated polyspectrum (bispectrum and trispectrum) based approaches are described in Section IV. Consistency of the proposed parameter estimators is established in Section V under some mild sufficient conditions. Finally, two simulation examples are presented in Section VI to illustrate the proposed approaches and to compare them with several existing approaches. Certain technical details may be found in the Appendix.

II. MODEL ASSUMPTIONS AND CUMULANT SPECTRA

Let $u(t)$ and $s(t)$ denote the "true" input and output, respectively, at (discrete) time t , related via a finite-dimensional ARMA(n_a, n_b) model

$$s(t) + \sum_{i=1}^{n_b} a_i s(t-i) = \sum_{i=1}^{n_a} b_i u(t-i). \quad (2-1)$$

As in [1] and [7], we have assumed for convenience only that no term $b_0 u(t)$ is present, i.e., there is at least one delay in the system. The processes $\{u(t)\}$ and $\{s(t)\}$ are not available for measurement. But we can measure noise-contaminated input and output

$$x(t) = u(t) + v_i(t) \quad (2-2)$$

$$y(t) = s(t) + v_o(t). \quad (2-3)$$

The model (2-1) is assumed to be exponentially stable and minimal, i.e., the following condition is assumed to be true.

H1) $A(z) = 1 + \sum_{i=1}^{n_b} a_i z^{-i} \neq 0$ for $|z| \geq 1$ where z is a complex variable. Moreover, $A(z)$ and $B(z) = \sum_{i=1}^{n_a} b_i z^{-i}$ are coprime (they have no common factors).

It is also assumed that all of the processes involved (i.e., $x(t)$, $y(t)$, $v_i(t)$, and $v_o(t)$) are zero-mean and jointly stationary. Furthermore, the noise sequences $\{v_i(t)\}$ and $\{v_o(t)\}$ are independent of $\{u(t)\}$, hence of $\{s(t)\}$.

Define the vector of the unknown parameters in the system model as

$$\theta(n_a, n_b) := (a_1, a_2, \dots, a_{n_b}, b_1, b_2, \dots, b_{n_a}). \quad (2-4)$$

Let θ_0 , n_{a0} and n_{b0} denote the true values of the respective vector/parameters. Also define the following set

$$\Theta(n_a, n_b) := \{\theta(n_a, n_b) | A(z; \theta(n_a, n_b)) \neq 0 \text{ for } |z| \geq 1\}$$

where $A(z; \theta(n_a, n_b))$ denotes $A(z)$ explicitly parameterized by $\theta(n_a, n_b)$. Where there is no risk of confusion, we will suppress explicit dependence upon the model orders, e.g., we will denote $\theta(n_a, n_b)$ by θ , $\Theta(n_a, n_b)$ by Θ , etc. The objective is to estimate the transfer function $B(z)/A(z)$ (equivalently, θ) given a data record $\{x(i), y(i), 1 \leq i \leq N\}$.

Consider the third-order cumulant function $C_{xy}(i, k)$ defined as

$$C_{xy}(i, k) := E\{x(t+i)x(t+k)y(t)\}. \quad (2-5)$$

Denote the cross-bispectrum of input/output by $B_{xy}(\omega_1, \omega_2)$, given by

$$B_{xy}(\omega_1, \omega_2) = \sum_{i=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_{xy}(i, k) \cdot \exp\{-j(\omega_1 i + \omega_2 k)\}. \quad (2-6)$$

Similarly, let $B_{xxx}(\omega_1, \omega_2)$ and $B_{sss}(\omega_1, \omega_2)$ denote the bispectra of the processes $\{x(t)\}$ and $\{s(t)\}$, respectively. From the above definitions it is easy to see that

$$C_{ssx}(i, k) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} B_{ssx}(\omega_1, \omega_2) \cdot \exp\{j(\omega_1 i + \omega_2 k)\} d\omega_1 d\omega_2. \quad (2-7)$$

Define

$$w(t) := s^2(t) - E\{s^2(t)\} \quad \text{and} \quad v(t) := s^2(t). \quad (2-8)$$

Then both the cross-spectrum between the process $\{w(t)\}$ and $\{s(t)\}$ and the cross-spectrum between the process $\{w(t)\}$ and $\{s(t)\}$ are given by

$$\begin{aligned} S_{ws}(\omega) &:= \sum_{k=-\infty}^{\infty} E\{w(t+k)s(t)\} \exp\{-j\omega k\} \\ &= \sum_{k=-\infty}^{\infty} C_{ws}(k, k) \exp\{-j\omega k\} \\ &= S_{ws}(\omega) \end{aligned} \quad (2-9)$$

It then follows that

$$\begin{aligned} C_{ws}(k, k) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{ws}(\omega) \exp\{j\omega k\} d\omega \\ &= C_{ss}(0, -k) = C_{ss}^*(0, -k) \end{aligned} \quad (2-10)$$

where $*$ denotes complex conjugation. Compare (2-7) with (2-10) to deduce that

$$\begin{aligned} S_{ws}^*(\omega) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} B_{ws}(\omega, \omega_2) d\omega_2 \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} B_{ss}(\omega_1, \omega) d\omega_1 \\ &= S_{ss}^*(\omega). \end{aligned} \quad (2-11)$$

Notice that the cross-spectrum between the signal $s(t)$ and its square can be interpreted as an integrated bispectrum of $s(t)$. This integrated bispectrum will form a basis (along with the integrated trispectrum, to be defined later) for unknown parameter estimation. It is easy to see that since bispectrum of a Gaussian process is identically zero, so is its integrated bispectrum. Therefore, the integrated bispectrum of $\{y(t)\}$ equals the integrated bispectrum of $\{s(t)\}$. In the sequel it will be easier to work with the centered (zero-mean) $s^2(t)$, i.e., $w(t)$.

Turning to the trispectrum, it is defined as ([23] and [24])

$$T_{ssss}(\omega_1, \omega_2, \omega_3) := \sum_{i=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} C_{ssss}(i, k, l) \cdot \exp\{-j(\omega_1 i + \omega_2 k + \omega_3 l)\} \quad (2-12)$$

where

$$\begin{aligned} C_{ssss}(i, k, l) &:= E\{s(t)s(t+i)s(t+k)s(t+l)\} \\ &\quad - E\{s(t)s(t+i)\}E\{s(t+k)s(t+l)\} \\ &\quad - E\{s(t)s(t+k)\}E\{s(t+l)s(t+i)\} \\ &\quad - E\{s(t)s(t+l)\}E\{s(t+k)s(t+i)\} \\ &\quad + E\{s(t)s(t+k)s(t+i)\}E\{s(t+l)\} \end{aligned} \quad (2-13)$$

is the fourth-order cumulant function of the process $\{s(t)\}$. It then follows that

$$\begin{aligned} C_{ssss}(i, k, l) &= \frac{1}{(2\pi)^3} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} T_{ssss}(\omega_1, \omega_2, \omega_3) \\ &\quad \cdot \exp\{j(\omega_1 i + \omega_2 k + \omega_3 l)\} \\ &\quad \cdot d\omega_1 d\omega_2 d\omega_3. \end{aligned} \quad (2-14)$$

Define

$$\tilde{r}(t) := s^3(t) - 3s(t)E\{s^2(t)\}$$

and

$$r(t) = \tilde{r}(t) - E\{\tilde{r}(t)\} \quad (2-15)$$

Then both the cross-spectrum between the process $\{r(t)\}$ and $\{s(t)\}$ and the cross-spectrum between the process $\{r(t)\}$ and $\{s(t)\}$ are given by

$$\begin{aligned} S_{rs}(\omega) &:= \sum_{k=-\infty}^{\infty} E\{r(t+k)s(t)\} \exp\{-j\omega k\} \\ &= \sum_{k=-\infty}^{\infty} C_{rs}(k, k) \exp\{-j\omega k\} \\ &= S_{rs}(\omega). \end{aligned} \quad (2-16)$$

It then follows that

$$\begin{aligned} C_{rs}(k, k, k) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{rs}(\omega) \exp\{j\omega k\} d\omega \\ &= C_{ssss}(0, 0, -k) \\ &= C_{ssss}^*(0, 0, -k). \end{aligned} \quad (2-17)$$

Compare (2-14) with (2-17) to deduce that

$$\begin{aligned} S_{rs}^*(\omega) &= \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} T_{ssss}(\omega, \omega_2, \omega_3) d\omega_2 d\omega_3 \\ &= S_{rs}^*(\omega). \end{aligned} \quad (2-18)$$

Notice that the cross-spectrum between the signal $s(t)$ and a function of its cube can be interpreted as an integrated trispectrum of $s(t)$.

Given the above definitions we are now ready to state the remaining model assumptions.

Assumption Set I: Assume that the bispectra of the noise processes $\{v_i(t)\}$ and $\{v_o(t)\}$ are zero and that the noise processes are statistically independent of the process $\{u(t)\}$, hence of $\{s(t)\}$. It is also assumed that $B_{uuu}(\omega_1, \omega_2) \neq 0$ and that H1) holds. Assume that all moments of the various processes involved, viz., $s(t)$, $u(t)$ etc., exist. Also assume that the third and lower order cumulant/cross-cumulant sequences of the various processes involved, viz., $C_{sux}(\tau_1, \tau_2)$ etc., satisfy the following summability conditions

$$\sum_{\tau_1, \dots, \tau_{k-1} = -\infty}^{\infty} [1 + |\tau_j|] |C_{z_1 z_2 \dots z_k}(\tau_1, \dots, \tau_{k-1})| < \infty$$

for each $j = 2, \dots, k-1$ and each $k = 2, 3, \dots$ where $z_i(t) \in \{s(t), u(t), x(t), y(t), v_i(t), v_o(t)\}$. The above summability conditions are sufficient for the corresponding bispectrum to exist [24], [42], [43]. They are also sufficient for the asymptotic results concerning the bispectrum estimator discussed in Section III-A-1) and concerning the integrated bispectrum discussed in Section IV-A to hold. \square

Assumption Set II: In addition to those stated in Assumption Set I, assume that $\{u(t)\}$ is such that at least one of the following conditions hold true:

AS1) $E\{u^3(t)\} = \gamma_{3u} \neq 0$.

AS2) $\{u(t)\}$ is a finite-dimensional, linear non-Gaussian process generated by driving a bounded-input bounded-output linear filter by a zero-mean, i.i.d. sequence $\{\epsilon(t)\}$

$$u(t) = \sum g(i)\epsilon(t-i) \quad (2-19)$$

where $\{g(i), -\infty < i < \infty\}$ is the impulse response of the filter and $E\{\epsilon^3(t)\} \neq 0$. \square

Assumption Set III: Assume that the trispectra of the noise processes $\{v_i(t)\}$ and $\{v_o(t)\}$ are zero and that the noise processes are statistically independent of the process $\{u(t)\}$, hence of $\{s(t)\}$. Also assume that $T_{uuuu}(\omega_1, \omega_2, \omega_3) \neq 0$. It is assumed that H1) holds and that all moments of the various processes involved, viz., $s(t)$, $u(t)$ etc., exist. Also assume that the fourth and lower order cumulant/cross-cumulant sequences of the various processes involved, viz., $C_{suxy}(\tau_1, \tau_2, \tau_3)$ etc., satisfy the following summability conditions

$$[1 + |\tau_j|] |C_{z_1 z_2 \dots z_k}(\tau_1, \dots, \tau_{k-1})| < \infty$$

for each $j = 1, 2, \dots, k-1$ and each $k = 2, 3, \dots$ where $z_i(t) \in \{s(t), u(t), x(t), y(t), v_i(t), v_o(t)\}$. The above summability conditions are sufficient for the corresponding trispectrum (or cross-trispectrum) to exist [24], [42], [43]. They are also sufficient for the asymptotic results concerning the integrated trispectrum discussed in Section IV-A to hold. Assume also that $\{u(t)\}$ is such that at least one of the following conditions hold true:

AS3) $\gamma_{4u} := E\{u^4(t)\} - 3[E\{u^2(t)\}]^2 \neq 0$.

AS4) $\{u(t)\}$ is a finite-dimensional, linear non-Gaussian process as in AS2) except that now $\gamma_{4\epsilon} \neq 0$. \square

Suppose that Assumption AS1) is true. Then we can not have $S_{u^2u}(\omega) \equiv 0$ because then $\gamma_{3u} = 0$ contrary to the assumption. Suppose that Assumption AS2) is true. Then using (2-19) it can be shown that [23] and [24] $S_{u^2u}(\omega) = \gamma_{3\epsilon} G_2(-\omega)G(\omega)$, where $\gamma_{3\epsilon} = E\{\epsilon^3(t)\}$, $G(\omega) = \sum_{k=-\infty}^{\infty} g(k) \exp\{-j\omega k\}$ and $G_2(\omega) = \sum_{k=-\infty}^{\infty} g^2(k) e^{-j\omega k}$. It then follows that unless $G(\omega) \equiv 0$, we have $S_{u^2u}(\omega) \neq 0$. Thus, under either AS1) or AS2) (or both), the integrated bispectrum of $u(t)$ is nonvanishing. Turning to the integrated trispectrum, it follows that under Assumption AS3) (i.e., $\gamma_{4u} \neq 0$), we must have $S_{r_{uu}}(\omega) \neq 0$ (where $r_u(t) := u^3(t) - 3u(t)E\{u^2(t)\}$), else we have a contradiction. Under AS4), $\gamma_{4u} = \gamma_{4\epsilon} \sum_k g^4(k) \neq 0$ unless $G(\omega) \equiv 0$ or $\gamma_{4\epsilon} = 0$.

III. BISPECTRUM BASED APPROACHES

In this section two new parametric frequency domain approaches are proposed for estimation of the parameters of linear errors-in-variables models. One of the proposed approaches is a linear estimator where using the bispectrum of the input and the cross-bispectrum of the input-output, the system transfer function is first estimated at a number of frequencies exceeding one-half the number of unknown parameters. The estimated transfer function is used to estimate the unknown parameters using an overdetermined linear system of equations. In the second approach a quadratic transfer function matching criterion is optimized by using the linear estimator as an initial guess. The input to the system need not be a linear process but must have nonvanishing bispectrum. The estimators are later analyzed in Section V under Assumption Set I.

A. Linear Estimator

It follows from (2-1)–(2-3), (2-5), and (2-6) that

$$B_{xy}(\omega_1, \omega_2) = H^*(e^{j(\omega_1 + \omega_2)}) B_{xx}(\omega_1, \omega_2)$$

where $H(z) = \sum_{i=1}^n b_i z^{-i} / [1 + \sum_{i=1}^n a_i z^{-i}]$ and H^* is the complex conjugate of H . Therefore, if $B_{xx}(\omega_1, \omega_2) \neq 0$ then

$$H^*(e^{j(\omega_1 + \omega_2)}) = B_{xy}(\omega_1, \omega_2) / B_{xx}(\omega_1, \omega_2). \quad (3-1)$$

The above relation occurs in [5] and [41]. Our new linear approach is to first rewrite (3-1) as

$$H(e^{j\omega}) = B_{xy}^*(\omega_1, \omega - \omega_1) / B_{xx}^*(\omega_1, \omega - \omega_1). \quad (3-2)$$

Thus, for a given ω several estimates of $H(e^{j\omega})$ can be obtained from (3-2) by using different values of ω_1 . This, in turn, can be used to devise parametric estimators for the model coefficients; details are in Sections III-A-2), III-B, and III-C.

1) **Estimation of Bispectrum and Cross-Bispectrum:** We now discuss estimation of the bispectra given data $\{x(t), y(t), 1 \leq t \leq N\}$ and some large sample properties of the estimators. Of the various approaches available in [21], [22], [25], and [30], we will follow the approach of [22] and [30]. Suppose that the given sample sequence of length N is divided into K nonoverlapping segments each of size L_B samples so that $N = KL_B$. Let $X^{(i)}(\omega)$

denote the discrete Fourier transform (DFT) of the i th block $\{x(t + (i-1)L_B), 1 \leq t \leq L_B\}$ ($i = 1, 2, \dots, K$) given by

$$X^{(i)}(\omega_k) = \sum_{t=0}^{L_B-1} x(t + 1 + (i-1)L_B) \exp(-j\omega_k t) \quad (3-3)$$

where

$$\omega_k = \frac{2\pi}{L_B} k, \quad k = 0, 1, \dots, L_B - 1. \quad (3-4)$$

Similarly define $Y^{(i)}(\omega_k)$. Then the bispectrum estimate $\hat{B}_{xxx}(m, n)$ at bifrequency (ω_m, ω_n) is given by averaging the "biperiodograms" over K blocks and spatially smoothing over a rectangular window of size $M \times M$ where $M = 2\bar{M} + 1$ is an odd integer

$$\begin{aligned} \hat{B}_{xxx}(m, n) = & \frac{1}{M^2} \sum_{r=-\bar{M}}^{\bar{M}} \sum_{s=-\bar{M}}^{\bar{M}} \left[\frac{1}{K} \sum_{i=1}^K \right. \\ & \cdot \left(\frac{1}{L_B} X^{(i)}(\omega_{m+r}) X^{(i)}(\omega_{n+s}) \right. \\ & \cdot \left. [X^{(i)}(\omega_{m+r} + \omega_{n+s})]^* \right) \end{aligned} \quad (3-5)$$

and similarly we have

$$\begin{aligned} \hat{B}_{xxy}(m, n) = & \frac{1}{M^2} \sum_{r=-\bar{M}}^{\bar{M}} \sum_{s=-\bar{M}}^{\bar{M}} \left[\frac{1}{K} \sum_{i=1}^K \right. \\ & \cdot \left(\frac{1}{L_B} X^{(i)}(\omega_{m+r}) X^{(i)}(\omega_{n+s}) \right. \\ & \cdot \left. [Y^{(i)}(\omega_{m+r} + \omega_{n+s})]^* \right) \end{aligned} \quad (3-6)$$

The principal domain (nonredundant region) of $\hat{B}_{xxx}(m, n)$ is the triangular grid

$$\begin{aligned} D_{xxx} = & \{(m, n) | m = Mk - \bar{M}, n = Ml - \bar{M}, \\ & 1 \leq l \leq k \leq \frac{L_B}{2M} - l, (m, n) \in \bar{D}_{xxx}\} \end{aligned} \quad (3-7)$$

where $\bar{D}_{xxx} = \{(k, l) | 0 \leq k \leq L_B/2, 0 \leq l \leq k, 2k + l \leq L_B\}$. The principal domain of $\hat{B}_{xxy}(m, n)$ is a larger triangular grid [41]

$$D_{xxy} = D_{xxy}^{(1)} \cup D_{xxy}^{(2)} \quad (3-8)$$

where $D_{xxy}^{(1)} = \{(m, n) | m = Mk - \bar{M}, n = Ml - \bar{M}, 1 \leq l \leq k \leq L_B/2M, (m, n) \in \bar{D}_{xxy}, n \geq 0\}$, $D_{xxy}^{(2)} = \{(m, n) | (m, -n) \in D_{xxy}^{(1)}\}$ and $\bar{D}_{xxy} = \{(k, l) | 0 \leq k \leq L_B/2, -L_B/2 < l \leq L_B/2, k \geq |l|\}$.

Define $\Delta_N = M/L_B = MK/N$. The quantity Δ_N is a measure of the "bandwidth" of the bispectrum estimate. Under the cumulant summability condition stated in Assumption Set I, it follows from [24, Chapter VII], [42, Chapter 4], and [43] that

$$E\{\hat{B}_{xxx}(m, n)\} = B_{xxx}(\omega_m, \omega_n) + O(\Delta_N), \quad (3-9)$$

$$\begin{aligned} \text{var}\{\text{Re}\{\hat{B}_{xxx}(m, n)\}\} = & \frac{1}{2N\Delta_N^2} S_x(\omega_m) S_x(\omega_n) \\ & \cdot S_x(\omega_{m+n}) + O\left(\frac{1}{N\Delta_N}\right) \\ = & \text{var}\{\text{Im}\{\hat{B}_{xxx}(m, n)\}\} \end{aligned} \quad (3-10)$$

where $S_x(\omega_m)$ denotes the power spectral density of $\{x(t)\}$ at frequency $\omega_m = 2\pi m/L_B$. Note that $N\Delta_N^2 = KM^2/L_B$. Therefore, as $N \rightarrow \infty$ such that $\Delta_N \rightarrow 0$ and $N\Delta_N^2 \rightarrow \infty$, we obtain an unbiased and mean-square consistent estimator. Moreover, under the above asymptotic conditions, the estimate $\hat{B}_{xxx}(m, n)$ ($m, n \neq 0$), is complex Gaussian (with asymptotically independent real and imaginary parts), and independent of $\hat{B}_{xxx}(m', n')$ ($m', n' \neq 0$) if $m' \neq m$ and/or $n' \neq n$, such that both (m, n) and (m', n') are in the interior of D_{xxx} . (See [42, Section 4.2] for a definition of complex Gaussian random variables.) Similar results holds for cross-bispectrum estimates leading to

$$E\{\hat{B}_{xxy}(m, n)\} = B_{xxy}(\omega_m, \omega_n) + O(\Delta_N), \quad (3-11)$$

$$\begin{aligned} \text{var}\{\text{Re}\{\hat{B}_{xxy}(m, n)\}\} = & \frac{1}{2N\Delta_N^2} S_x(\omega_m) S_x(\omega_n) \\ & S_y(\omega_{m+n}) + O\left(\frac{1}{N\Delta_N}\right) \\ = & \text{var}\{\text{Im}\{\hat{B}_{xxy}(m, n)\}\}. \end{aligned} \quad (3-12)$$

The estimate $\hat{B}_{xxy}(m, n)$ ($m, n \neq 0$) is complex Gaussian (with asymptotically independent real and imaginary parts) and independent of $\hat{B}_{xxy}(m', n')$ ($m', n' \neq 0$) if $m' \neq m$ and/or $n' \neq n$, such that both (m, n) and (m', n') are in the interior of D_{xxy} . As noted before following (3-10), as $N \rightarrow \infty$ such that $\Delta_N \rightarrow 0$ and $N\Delta_N^2 \rightarrow \infty$, we obtain an unbiased and mean-square consistent estimator. One way to accomplish this is to choose $\Delta_N = N^{c-0.5}$ with $0 < c < 0.5$. In terms of K , M , and L_B , one may pick $K = 1$, $L_B = N$, and $M = N^{c+0.5}$ with $0 < c < 0.5$. Alternatively, one may choose $M = 1$, $L_B = N^{0.5-c}$, and $K = N^{0.5+c}$ with $0 < c < 0.5$. The smallest bias results with $c \rightarrow 0$ whereas the smallest variance occurs for $c \rightarrow 0.5$.

2) *Overdetermined Linear System of Equations:* Returning to (3-2) and using the estimates of the bispectrum and cross-bispectrum, we have

$$\begin{aligned} \hat{H}(\omega_n; \omega_m) := & \frac{\hat{B}_{xxy}(m, n-m)}{\hat{B}_{xxx}(m, n-m)} \quad \text{for } (m, n-m) \in D_{xxy} \\ \text{and} \end{aligned}$$

$$n = lM - \bar{M}, 1 \leq l \leq \frac{L_B}{2M} - 1.$$

Let $\hat{H}_B(e^{j\omega_n})$ denote the least-squares estimate of $H(e^{j\omega_n})$ given $\hat{H}(\omega_n; \omega_m)$ for a fixed n with m ranging over $(m, n-m) \in D_{xxy}$. Then we have

$$\begin{aligned} \hat{H}_B(e^{j\omega_n}) = & \frac{\hat{B}_{xxx}(m, n-m) \hat{B}_{xxy}^*}{\sum_{(m, n-m) \in D_{xxy}} |\hat{B}_{xxx}(m, n-m)|^2} \quad (3-13) \end{aligned}$$

for $n = lM - \bar{M}$, $1 \leq l \leq (L_B/(2M)) - 1$.

From the definition of $H(e^{j\omega})$ we have

$$-\sum_{i=1}^{n_a} a_i H(e^{j\omega_n}) e^{-j\omega_n i} + \sum_{i=1}^{n_b} b_i e^{-j\omega_n i} = H(e^{j\omega_n}) \quad (3-14)$$

for any ω_n . Noting that a_i 's and b_i 's are real and $H(e^{j\omega})$ is, in general, complex-valued, we rewrite (3-14) after replacing $H(e^{j\omega_n})$ with its least-square estimate $\hat{H}_B(e^{j\omega_n})$, as

$$-\sum_{i=1}^{n_a} a_i \operatorname{Re} \{ \hat{H}_B(e^{j\omega_n}) e^{-j\omega_n i} \} + \sum_{i=1}^{n_b} b_i \operatorname{Re} \{ e^{-j\omega_n i} \} = \operatorname{Re} \{ \hat{H}_B(e^{j\omega_n}) \} \quad (3-15)$$

and

$$-\sum_{i=1}^{n_a} a_i \operatorname{Im} \{ \hat{H}_B(e^{j\omega_n}) e^{-j\omega_n i} \} + \sum_{i=1}^{n_b} b_i \operatorname{Im} \{ e^{-j\omega_n i} \} = \operatorname{Im} \{ \hat{H}_B(e^{j\omega_n}) \}. \quad (3-16)$$

We solve (3-15) and (3-16) for a_i 's and b_i 's with $n = lM - \bar{M}$, $1 \leq l \leq L_B/2M - 1$ using an ordinary linear least-squares formulation.

B. Nonlinear Estimator

We follow a quadratic transfer function matching approach. Let $H(e^{j\omega}|\theta)$ denote the transfer function of (2-1) with the system parameters specified by the parameter vector θ as defined in (2-4). Define

$$H(e^{j(\omega_m + \omega_n)}) = \frac{B_{xy}^*(m, n)}{\hat{B}_{xx}^*(m, n)}. \quad (3-17)$$

The following result is from [24] and [43]. A proof of Lemma 1 also follows from the mean-square convergence implied by (3-9)–(3-12).

Lemma 1: Under the Assumption Set 1, we have $\lim_{N \rightarrow \infty} B_{xx}(m, n) = B_{xx}(\omega_m, \omega_n)$ i.p. (in probability), and $\lim_{N \rightarrow \infty} B_{xy}(m, n) = B_{xy}(\omega_m, \omega_n)$ i.p. •

Using Lemma 1 and the asymptotic distribution of the bispectrum estimators, the following useful result can be established.

Lemma 2: Under the Assumption Set 1 as $N \rightarrow \infty$, the following results are true for any fixed (ω_m, ω_n) and $(\omega_{m'}, \omega_{n'})$ in the interior of the set D_{xy} .

- A) $\Delta_N \sqrt{N} (H(e^{j(\omega_m + \omega_n)}) - H(e^{j(\omega_m + \omega_n)}|\theta_0))$ converges in distribution to the complex normal distribution $\mathcal{N}_c(0, \sigma_{mn}^2)$ where

$$\sigma_{mn}^2 = \frac{g(m, n) S_x(\omega_m) S_x(\omega_n) S_y(\omega_{m+n})}{|B_{xx}(\omega_m, \omega_n)|^2}.$$

$$g(m, n) = 1 + |\beta_{m, n}|^2 \frac{S_{xx}(\omega_{m+n})}{S_{yy}(\omega_{m+n})} - 2 \operatorname{Re} \left\{ \beta_{m, n} \frac{S_{xy}(\omega_{m+n})}{S_{yy}(\omega_{m+n})} \right\}$$

and $\beta_{m, n} = B_{xy}^*(\omega_m, \omega_n) / B_{xx}^*(\omega_m, \omega_n)$.

- B) $\hat{H}(e^{j(\omega_m + \omega_n)})$ and $\hat{H}(e^{j(\omega_{m'} + \omega_{n'})})$ are statistically independent for $(m, n) \neq (m', n')$. •

Sketch of Proof: It follows from the asymptotic results (3-7)–(3-12) discussed in Section III-A-1, Lemma 1, Slutsky's Theorem [36, Section 4.1], and [42, Theorem P5.2]. □

Lemma 2 motivates the following cost criterion which is the negative log-likelihood (up to some constants which do not depend upon θ) of the asymptotically complex Gaussian vector $\{\hat{H}(e^{j(\omega_m + \omega_n)})\}$, $(m, n) \in D_{xy}^0$ where D_{xy}^0 is the interior of D_{xy} . Choose θ to minimize the cost

$$\sum_{(m, n) \in D_{xy}^0} \frac{|\hat{B}_{xy}^*(m, n) - H(e^{j(\omega_m + \omega_n)}|\theta)|^2}{\sigma_{mn}^2} \quad (3-18)$$

In the above we have assumed that $\sigma_{mn}^2 \neq 0$. If this is not the case then we delete that particular value of (m, n) from the summation. In practice, $S_x(\omega_m)$ and $S_y(\omega_m)$ are unknown; therefore, we replace them by their consistent estimates obtained by block averaging and/or frequency-domain smoothing as in (3-5)–(3-5). Thus we have the following practical cost criterion

$$J_N^{(B)}(\theta) = \sum_{(m, n) \in D_{xy}^0} \frac{|\hat{B}_{xy}^*(m, n) - H(e^{j(\omega_m + \omega_n)}|\theta)|^2}{\hat{\sigma}_{mn}^2} \quad (3-19)$$

where

$$\hat{\sigma}_{mn}^2 = \hat{g}(m, n) \hat{S}_x(m) \hat{S}_x(n) \cdot S_y(m+n) / |\hat{B}_{xx}(m, n)|^2 \quad (3-20)$$

and $\hat{g}(m, n)$ follows similarly by replacing the variables in $g(m, n)$ with their consistent estimates. Mimicking Lemma 1 and using Slutsky's Theorem, it is easy to show that $\hat{\sigma}_{mn}^2$ tends to σ_{mn}^2 i.p. as $N \rightarrow \infty$ provided $B_{xx}(\omega_m, \omega_n) \neq 0$. Minimization of (3-19) w.r.t. θ is a nonlinear optimization problem that is initialized by the linear estimator obtained by solving (3-15) and (3-16).

C. Summary of Algorithms

Summarizing the above two algorithms, we fit an ARMA(n_a, n_b) model as follows.

- Given the record length N , select the block length L_B , number of blocks K , and the smoothing window size M . See the discussion in Section III-A-1 [following (3-12)] for possible choices.
- Estimate the required bispectra using (3-5) and (3-6), and the spectra \hat{S}_{xx} and \hat{S}_{yy} using similar equations.
- Estimate the model coefficients via an ordinary least-squares formulation using (3-13), (3-15)–(3-16). (Use all possible bifrequencies in (3-13) unless $|\hat{B}_{xx}(m, n - m)|$ is too small. Since the bispectrum estimates at distinct bifrequencies are approximately statistically independent, the estimates at distinct bifrequencies carry “new” information.) Stop if desired only linear estimators.

- iv) Using the estimate obtained in iii) as an initial guess, refine the model coefficient estimates by minimizing (3-19). Again, if possible, use all possible bifrequencies.

IV. INTEGRATED POLYSPECTRUM BASED APPROACHES

In this section, the frequency domain approaches presented in Section III are discussed using integrated polyspectra. As discussed in Section II, integrated polyspectrum (bispectrum or trispectrum) is computed as a cross-spectrum; hence, it is computationally cheaper than the corresponding polyspectrum particularly in the case of the fourth-order cumulant spectrum. If the non-Gaussian input to the linear system has a symmetric PDF, its bispectrum and the integrated bispectrum will vanish whereas its trispectrum and the integrated trispectrum will not, provided that the fourth cumulant of the input γ_{4u} is nonzero. Extension of the approaches of Section III to trispectrum-based approaches is computationally complex and the resulting estimators are likely to have poor statistical performance because of the high variance of the trispectrum estimators [24]. Herein lies the significance of the integrated polyspectrum-based approaches which apply with almost equal computational and programming ease to both cases, those involving integrated bispectrum as well as those concerned with integrated trispectrum.

A. Linear Estimator

Define

$$r_{2x}(t) = x^2(t) \quad (4-1)$$

and

$$r_{3x}(t) = x^3(t) - 3x(t)E\{x^2(t)\} - E\{x^3(t)\} \quad (4-2)$$

Replacing $x(t)$ with $u(t)$ in (4-2) yields $r_{3u}(t)$ which is identical to $r_{uu}(t)$ in Section II. In the following we will use the notation $r_{\tau}(t)$ to refer to both $r_{2x}(t)$ and $r_{3x}(t)$, meaning (4-1) when discussing integrated bispectrum and implying (4-2) when discussing integrated trispectrum.

Under the Assumption Set II (or III), we have

$$S_{r_{xx}}(\omega) = S_{r_{uu}}(\omega) = \sum_{k=-\infty}^{\infty} E\{r_x(t+k)y(t)\} \exp(-j\omega k).$$

It follows from (2-1)–(2-3) that under the Assumption Set II (or III)

$$S_{r_{xy}}(\omega) = H^*(e^{j\omega})S_{r_{xx}}(\omega)$$

where $H(z) = B(z)/A(z)$, $S_{r_{xx}}(\omega) = S_{r_{uu}}(\omega)$ and $S_{r_{xy}}(\omega) = S_{r_{yu}}(\omega)$. Therefore, if $S_{r_{xx}}(\omega) \neq 0$ then

$$H^*(e^{j\omega}) = S_{r_{xy}}(\omega)[S_{r_{xx}}(\omega)]^{-1}. \quad (4-3)$$

1) Estimation of Integrated Auto- and Cross-Polyspectrum:

We now discuss estimation of the integrated polyspectra given data $\{x(t), y(t), 1 \leq t \leq N\}$ and some large sample properties of the estimators. Suppose that the given sample sequence of length N is divided into K nonoverlapping segments each of size L_B samples so that $N = KL_B$. Let $X^{(i)}(\omega)$ denote the DFT of the i th block $\{x(t+(i-1)L_B), 1 \leq t \leq L_B\}$ ($i = 1, 2, \dots, K$) as in (3-3) and (3-4). Similarly let $R_x^{(i)}(\omega)$ denote the DFT of the i th block of $r_x(t)$ given by

$$R^{(i)}(\omega_k) = \sum_{t=0}^{L_B-1} [x^3(t+1+(i-1)L_B) - 3x(t+1+(i-1)L_B)\hat{\mu}_{2x}] \cdot \exp(-j\omega_k t), \quad (4-4)$$

$$\hat{\mu}_{2x} = \frac{1}{N} \sum_{t=1}^N x^2(t) \quad (4-5)$$

when $r_x(t) = r_{3x}(t)$ [see (4-2)], and by

$$R^{(i)}(\omega_k) = \sum_{t=0}^{L_B-1} x^2(t+1+(i-1)L_B) \exp(-j\omega_k t) \quad (4-6)$$

when $r_x(t) = r_{2x}(t)$ [see (4-1)]. To set the mean of $\{r_x(t)\}$ in the i th block to zero, simply set $R^{(i)}(0) = 0$. Then the estimate $\hat{S}_{r_x}(\omega_m)$ of $S_{r_x}(\omega_m)$ based upon the record $\{x(t), y(t), 1 \leq t \leq N\}$ is given by averaging the block periodograms over K blocks and spatially smoothing over a rectangular window of size M where $M = 2\bar{M} + 1$ is an odd integer

$$\hat{S}_{r_x}(\omega_m) = M^{-1} \sum_{l=-\bar{M}}^{\bar{M}} \left\{ \frac{1}{K} \sum_{i=1}^K \overline{R^{(i)}(\omega_{m+l})} \cdot R^{(i)}(\omega_{m+l}) \right\} \quad (4-7)$$

where $m = Ml - \bar{M}$, $1 \leq l \leq L_B/2M - 1$. The estimate $\hat{S}_{r_y}(\omega_m)$ is given similarly by

$$\hat{S}_{r_y}(\omega_m) = M^{-1} \sum_{l=-\bar{M}}^{\bar{M}} \left\{ \frac{1}{K} \sum_{i=1}^K \left[\frac{1}{L_B} R^{(i)}(\omega_{m+l}) \cdot [Y^{(i)}(\omega_{m+l})]^* \right] \right\} \quad (4-8)$$

where $m = Ml - \bar{M}$, $1 \leq l \leq L_B/2M - 1$.

Under the cumulant summability conditions stated in Assumption Set II (or III), it follows from [24, Chapter VII], [42, Problem 7.10.8], and [43] that for large N , the real and the imaginary parts of the estimate $\hat{S}_{r_{xy}}(\omega_m)$ ($m \neq 0$) are bivariate Gaussian, and $\hat{S}_{r_{xy}}(\omega_m)$ is independent of $\hat{S}_{r_{xy}}(\omega_n)$ ($n \neq 0$) for $m \neq n$ ($m, n = Ml - \bar{M}$, $1 \leq l \leq L_B/2M - 1$) such that

$$\begin{aligned} E\{\hat{S}_{r_{xy}}(\omega_m)\} &= S_{r_{xy}}(\omega_m) + O(\Delta_N), \\ \text{var}\{\text{Re}\{\hat{S}_{r_{xy}}(\omega_m)\}\} &= \frac{1}{2N\Delta_N} [S_{r_{xx}}(\omega_m)S_{yy}(\omega_m) \\ &\quad + \text{Re}\{S_{r_{xy}}^2(\omega_m)\}] + O(N^{-1}), \end{aligned} \quad (4-9)$$

$$\begin{aligned} & \text{cov} \{ \text{Re} \{ \hat{S}_{r_x y}(m) \}, \text{Im} \{ \hat{S}_{r_x y}(m) \} \} \\ &= \frac{1}{2N\Delta_N} \text{Im} \{ S_{r_x y}^2(\omega_m) \} + O(N^{-1}). \end{aligned} \quad (4-10)$$

In the above, as in Section III-A-1), $\Delta_N = M/L_B$ is the bandwidth of the cross-spectrum estimates. Moreover, as in Section III-A-1), we must have $\Delta_N \rightarrow 0$ and $N\Delta_N \rightarrow \infty$ as $N \rightarrow \infty$ to obtain an unbiased and mean-square consistent estimator. Similar results hold for $\hat{S}_{r_x x}(m)$.

2) *Overdetermined Linear System of Equations:* Returning to (4-3) we set

$$\hat{H}_{IP}(e^{j\omega_n}) = \hat{S}_{r_x y}^*(n) [\hat{S}_{r_x x}^*(n)]^{-1}. \quad (4-11)$$

Mimicking Section III-A-2), we solve (3-15) and (3-16) for a_l 's and b_l 's with $n = lM - \bar{M}$, $1 \leq l \leq L_B/2M - 1$ and using the above estimate of $\hat{H}_{IP}(e^{j\omega_n})$.

B. Nonlinear Estimator

We follow a quadratic transfer function matching approach. Let $H(e^{j\omega}|\theta)$ be as in Section III-B. Lemmas 3 and 4 are counterparts of Lemmas 1 and 2, respectively.

Lemma 3: Under the Assumption Set II (or III), we have $\lim_{N \rightarrow \infty} \hat{S}_{r_x x}(m) = S_{r_x x}(\omega_m)$ i.p. and $\lim_{N \rightarrow \infty} \hat{S}_{r_x y}(m) = S_{r_x y}(\omega_m)$ i.p. •

Lemma 4: Under the Assumption Set II (or III) as $N \rightarrow \infty$, the following results are true for any fixed $0 < \omega_m, \omega_{m'} < \pi$.

A) Let $H'_m = \sqrt{N\Delta_N} [\hat{H}_{IP}(e^{j\omega_m}) - H(e^{j\omega_m}|\theta_0)]$. Then $[\text{Re } H'_m \text{ Im } H'_m]^T$ converges in distribution to a real bivariate normal distribution $\mathcal{N}_2(0, \Sigma)$ where (Σ_{ij}) denotes the ij -element of matrix Σ) $\Sigma_{11} = \Gamma_1 + \text{Re } \Gamma_2$, $\Sigma_{22} = \Gamma_1 - \text{Re } \Gamma_2$, $\Sigma_{12} = \text{Im } \Gamma_2$

$$\begin{aligned} \Gamma_1 &= 0.5d(m)S_{r_x x}(\omega_m)S_{yy}(\omega_m)|S_{r_x x}(\omega_m)|^{-2}, \\ d(m) &= 1 + |\alpha_m|^2 \frac{S_{xx}(\omega_m)}{S_{yy}(\omega_m)} - 2\text{Re} \left\{ \alpha_m \frac{S_{xy}(\omega_m)}{S_{yy}(\omega_m)} \right\}, \\ \Gamma_2 &= 0 \end{aligned}$$

$$\text{and } \alpha_m = S_{r_x y}^*(\omega_m)/S_{r_x x}^*(\omega_m)$$

B) $\hat{H}_{IP}(e^{j\omega_m})$ and $\hat{H}_{IP}(e^{j\omega_{m'}})$ are statistically independent for $m \neq m'$. •

The negative log-likelihood of the asymptotically complex Gaussian vector $\{\hat{H}_{IP}(e^{j\omega_m}), m \in D_{r_x}\}$ where $D_{r_x} := \{m|m = lM - \bar{M}, 1 \leq l \leq (L_B/(2M)) - 1\}$, is similar to that in Section III-B. As in Section III-B, we choose θ to minimize the cost (cf. (3-19))

$$J_N^{(IP)}(\theta) = \sum_{m \in D_{r_x}} \frac{S_{r_x y}^*(m)}{S_{r_x x}^*(m)} - H(e^{j\omega_m}|\theta) \bigg/ \hat{\sigma}_m^2 \quad (4-12)$$

where

$$\hat{\sigma}_m^2 = \hat{d}(m)\hat{S}_{r_x x}(m)\hat{S}_{yy}(m)/|\hat{S}_{r_x x}(m)|^2 \quad (4-13)$$

and $\hat{d}(m)$ follows similarly by replacing the variables in $d(m)$ with their consistent estimates. As in Section III-B, it is easy to show that $\hat{\sigma}_m^2$ tends to σ_m^2 i.p. as $N \rightarrow \infty$ provided $S_{r_x x}(\omega_m) \neq 0$. Minimization of (4-12) w.r.t. θ is a nonlinear optimization problem that is initialized by the linear estimator of Section IV-A.

C. Summary of Algorithms

Summarizing the above two algorithms, we fit an ARMA(n_a, n_b) model as follows.

- i) This is same as item i) of Section III-C.
- ii) Estimate the required integrated polyspectra using (4-7) and (4-8), and the spectra $\hat{S}_{r_x x}$, $\hat{S}_{r_x y}$ and \hat{S}_{yy} using similar equations.
- iii) Estimate the model coefficients via an ordinary least-squares formulation using (4-11), (3-15), and (3-16) with $\hat{H}_B(\cdot)$ replaced with $\hat{H}_{IP}(\cdot)$. (Use all possible frequencies in (3-15) and (3-16) unless $|\hat{S}_{r_x x}(m)|$ is too small. Since the spectrum estimates at distinct frequencies are approximately statistically independent, the estimates at distinct frequencies carry "new" information.) Stop if desire only linear estimators.
- iv) Using the estimate obtained in iii) as an initial guess, refine the model coefficient estimates by minimizing (4-12). Again, if possible, use all possible frequencies as in iii) above.

V. CONSISTENCY

In this section we discuss the (weak) consistency of the proposed approaches (as $N \rightarrow \infty$). Let $H(e^{j\omega}; \theta_0)$ denote the transfer function of the true model (2-1) with $n_a = n_{a0}$, $n_b = n_{b0}$ and $\theta = \theta_0$. Consider the system of equations (3-14) rewritten as follows after replacing $H(e^{j\omega})$ with $H(e^{j\omega}; \theta_0)$

$$\begin{aligned} & - \sum_{i=1}^{n_a} a_i H(e^{j\omega_l}; \theta_0) e^{-j\omega_l i} + \sum_{i=1}^{n_b} b_i e^{-j\omega_l i} \\ &= H(e^{j\omega_l}; \theta_0), \quad 1 \leq l \leq L. \end{aligned} \quad (5-1)$$

Note that in (5-1) ω_l represents an arbitrary frequency, not necessarily equal to $2\pi l/L_B$.

The following result is central to the consistency results.

Lemma 5: Given the transfer function $H(e^{j\omega}; \theta_0)$ of model (2-1) at frequencies $0 < \omega_1 < \omega_2 < \dots < \omega_L < \pi$ such that $n_a + n_b \leq 2L$ and $\min(n_a - n_{a0}, n_b - n_{b0}) \geq 0$. Then the set of solutions to (5-1) is such that $H(e^{j\omega}; \theta) \equiv H(e^{j\omega}; \theta_0)$. •

Proof: If (5-1) is satisfied for some θ (defined in (2-4)), then $H(e^{j\omega_l}; \theta) = H(e^{j\omega_l}; \theta_0)$ for $1 \leq l \leq L$. Define

$$\begin{aligned} \bar{A}(z; \theta, \theta_0) &= A(z; \theta_0)B(z; \theta) \\ &= \left[\sum_{i=0}^{n_{a0}} a_i(\theta_0)z^{-i} \mid \sum_{i=1}^{n_b} b_i(\theta)z^{-i} \right] \\ &= \sum_{i=1}^{n_{a0}+n_b} \bar{a}_i(\theta, \theta_0)z^{-i} \end{aligned} \quad (5-2)$$

where $a_0(\theta_0) := 1$ so that $\bar{a}_1(\theta, \theta_0) = b_1(\theta)$. Also define

$$\begin{aligned} \bar{B}(z; \theta, \theta_0) &= A(z; \theta)B(z; \theta_0) \\ &= \left[\sum_{i=0}^{n_a} a_i(\theta)z^{-i} \mid \sum_{i=1}^{n_{b0}} b_i(\theta_0)z^{-i} \right] \\ &= \sum_{i=1}^{n_a+n_{b0}} \bar{b}_i(\theta, \theta_0)z^{-i} \end{aligned} \quad (5-3)$$

where $a_0(\theta) := 1$ so that $\bar{b}_1(\theta, \theta_0) = b_1(\theta_0)$. Clearly if $H(e^{j\omega_l}; \theta) = H(e^{j\omega_l}; \theta_0)$ for some l , then

$$\bar{A}(e^{j\omega_l}; \theta, \theta_0) = \bar{B}(e^{j\omega_l}; \theta, \theta_0) \quad (5-4)$$

since $H(e^{j\omega}; \theta) = B(e^{j\omega}; \theta)A^{-1}(e^{j\omega}; \theta)$ for any ω and θ . Using (5-2) and (5-3) this in turn implies that

$$\sum_{i=1}^{\bar{n}} [\bar{a}_i(\theta, \theta_0) - \bar{b}_i(\theta, \theta_0)] e^{-j\omega_l i} = 0 \quad (5-5)$$

where $\bar{n} = \max(n_{a0} + n_b, n_a + n_{b0})$ and we define $\bar{a}_i(\theta, \theta_0) = 0$ for $n_{a0} + n_b + 1 \leq i \leq \bar{n}$ and $\bar{b}_i(\theta, \theta_0) = 0$ for $n_a + n_{b0} + 1 \leq i \leq \bar{n}$. Since the coefficients $\bar{a}_i(\theta, \theta_0)$'s and $\bar{b}_i(\theta, \theta_0)$'s are real-valued, it also follows from (5-5) that

$$\sum_{i=1}^{\bar{n}} [\bar{a}_i(\theta, \theta_0) - \bar{b}_i(\theta, \theta_0)] e^{j\omega_l i} = 0. \quad (5-6)$$

By the hypothesis of the lemma, (5-5) and (5-6) are true for $1 \leq l \leq L$. This implies that the $(\bar{n} - 1)$ -degree polynomial equation

$$\sum_{i=0}^{\bar{n}-1} [\bar{a}_{i+1}(\theta, \theta_0) - \bar{b}_{i+1}(\theta, \theta_0)] z^i = 0 \quad (5-7)$$

has $2L$ roots. Under the hypothesis of the lemma, $2L \geq n_a + n_b \geq \bar{n}$ because $\min(n_a - n_{a0}, n_b - n_{b0}) \geq 0$. Now (5-7) can have only $\bar{n} - 1 < 2L$ roots; hence, we must have

$$\bar{a}_{i+1}(\theta, \theta_0) = \bar{b}_{i+1}(\theta, \theta_0) \quad \text{for } 0 \leq i \leq \bar{n} - 1 \quad (5-8)$$

leading to $H(e^{j\omega}; \theta) \equiv H(e^{j\omega}; \theta_0)$. \square

A. Bispectrum Based Approaches

Lemma 5 combined with the results of Theorems 1 and 2 (discussed later) motivate the following definition of persistence of excitation of inputs to linear systems for the purpose of consistent parameter estimation using the bispectrum. This is completely analogous to the concept of persistent excitation when using second-order statistics (see [7, Section 5.4]). Similar to the second-order case, if the input is persistently exciting as defined below, the matrix F_∞ defined in the proof of Theorem 1 will have full rank, leading to parameter identifiability.

Definition 1: A stationary input $\{u(t)\}$ is persistently exciting of order L w.r.t. the bispectral statistics if its bispectrum $B_{uuu}(\omega_m, \omega_n)$ is nonzero at L distinct bifrequencies in the interior of its principal domain.

Theorem 1: The linear parameter estimator discussed in Section III-A is (weakly) consistent under the Assumption Set I if (3-15) and (3-16) are solved at L frequencies $0 < \omega_1 < \omega_2 < \dots < \omega_L < \pi$ such that $n_a + n_b \leq 2L$, $\min(n_a - n_{a0}, n_b - n_{b0}) = 0$, and $B_{uuu}(\omega_m, \omega_{n-m}) \neq 0$ for all bifrequencies used in (3-13).

Proof: Equations (3-15) and (3-16) can be expressed as a linear system of equations

$$F_N \theta = f_N \quad (5-9)$$

where θ is the $(n_a + n_b)$ -column vector of unknown parameters, the matrix F_N is $2L \times (n_a + n_b)$ consisting of appropriate parts of the left side of (3-15) and (3-16), and f_N is an $2L$ -column vector consisting of the right side of (3-15) and (3-16). By (3-13), Lemma 1, and Slutsky's Theorem

$$\lim_{N \rightarrow \infty} \hat{H}_B(e^{j\omega_n}) = H(e^{j\omega_n}; \theta_0) \text{ i.p.} \quad (5-10)$$

Let F_∞ and f_∞ denote F_N and f_N , respectively, when the transfer function estimates are replaced with their true values. Then by Slutsky's Theorem we have

$$\lim_{N \rightarrow \infty} F_N = F_\infty \text{ i.p.} \quad \text{and} \quad \lim_{N \rightarrow \infty} f_N = f_\infty \text{ i.p.} \quad (5-11)$$

Since $\min(n_a - n_{a0}, n_b - n_{b0}) = 0$, by Lemma 5, (3-15) and (3-16) have a unique solution given by $\theta = \theta_0$ provided that $\hat{H}_B(e^{j\omega_n}) = H(e^{j\omega_n}; \theta_0)$ in (3-15) and (3-16). Hence, F_∞ has full rank so that $[F_\infty^T F_\infty]^{-1}$ exists; therefore, by Slutsky's Theorem and (5-11), $\lim_{N \rightarrow \infty} [F_N^T F_N]^{-1} = [F_\infty^T F_\infty]^{-1}$ i.p. Similar arguments lead to

$$\begin{aligned} \hat{\theta}_N &:= [F_N^T F_N]^{-1} F_N^T f_N \\ \hat{\theta}_0 &:= [F_\infty^T F_\infty]^{-1} F_\infty^T f_\infty \quad \text{as } N \rightarrow \infty. \end{aligned} \quad (5-12)$$

This proves the desired result. \square

Define

$$\hat{\theta}_N := \arg \left\{ \inf_{\theta \in \Theta_c} J_N^{(B)}(\theta) \right\} \quad (5-13)$$

where $J_N^{(B)}(\theta)$ is given by (3-19) and $\Theta_c \subset \Theta$ is a compact set. Convergence of this nonlinear estimator is discussed next.

Theorem 2: Suppose that Assumption Set I holds and $J_N^{(B)}(\theta)$ utilizes at least L distinct bifrequencies (ω_m, ω_n) in the interior of the principal domain of $B_{uuu}(\omega_m, \omega_n)$ such that $n_a + n_b \leq 2L$, and $B_{uuu}(\omega_m, \omega_n) \neq 0$ for all bifrequencies used in (3-19). If $\theta_0 \in \Theta_c$, then $\hat{\theta}_N$ defined in (5-13) converges i.p. to a set D_0 as $N \rightarrow \infty$ where

$$D_0 = \left\{ \theta \mid \frac{B(z; \theta)}{A(z; \theta)} = \frac{B(z; \theta_0)}{A(z; \theta_0)} \right\}.$$

Proof: See the Appendix. \square

The following corollary follows from Theorem 2 and some standard results (see, e.g., [7, Section 6.3]). The additional condition imposed in Corollary 1 ensures that the set D_0 consists of a single element θ_0 .

Corollary 1: Under the hypotheses of Theorem 2, if $\min(n_a - n_{a0}, n_b - n_{b0}) = 0$, then $\hat{\theta}_N$ converges i.p. to θ_0 as $N \rightarrow \infty$.

B. Integrated Polyspectrum Based Approaches

The following results mimic the corresponding results of Section V-A with obvious modifications. Therefore, we will simply state the main results without any proofs.

Definition 2: A stationary input $\{u(t)\}$ is persistently exciting of order L w.r.t. the integrated polyspectral statistics if its integrated polyspectrum $S_{r_{uu}}(\omega)$ is nonzero at L distinct frequencies in the interval $(0, \pi)$. •

Theorem 3: The integrated polyspectrum-based linear parameter estimator discussed in Section IV-A is (weakly) consistent under the Assumption Set II (or III) if (3-15) and (3-16) are solved (after replacing $\hat{H}_B(e^{j\omega_n})$ with $\hat{H}_{IP}(e^{j\omega_n})$) at L frequencies $0 < \omega_1 < \omega_2 < \dots < \omega_L < \pi$ such that $n_a + n_b \leq 2L$, $\min(n_a - n_{a0}, n_b - n_{b0}) = 0$, and $S_{r_{uu}}(\omega_m) \neq 0$ for all frequencies used in (3-15) and (3-16). •

Theorem 4: Consider the nonlinear parameter estimator defined by

$$\hat{\theta}_N := \arg \left\{ \inf_{\theta \in \Theta_C} J_N^{(IP)}(\theta) \right\}$$

where $\Theta_C \subset \Theta$ is a compact set. Suppose that Assumption Set II (or III) holds and $J_N^{(IP)}(\theta)$ utilizes at least L distinct frequencies ω_m in the interval $(0, \pi)$ such that $n_a + n_b \leq 2L$, $\min(n_a - n_{a0}, n_b - n_{b0}) = 0$, and $S_{r_{uu}}(\omega_m) \neq 0$ for all frequencies used in (4-12). If $\theta_0 \in \Theta_C$, then $\hat{\theta}_N$ converges i.p. to θ_0 as $N \rightarrow \infty$. •

VI. SIMULATION EXAMPLES

We now present two simulation examples to illustrate the proposed approaches. Example 1 is from [39], and Example 2 is from [34] and [40]. In both the examples the proposed approaches were applied using a block length $L_B = 64$ and a smoothing window $M = 1$ ($\bar{M} = 0$, i.e., no smoothing). The software package NL2SOL [15] was used for nonlinear optimization using the results of the linear approaches as initial guesses. Also the nonlinear cost functions were simplified somewhat by setting $\hat{g}(m, n) \equiv 1$ in (3-20) and $\hat{d}(m) \equiv 1$ in (4-13); this has no effect on the consistency results but will influence the accuracy of the estimates.

A. Example 1 [39]

The system to be identified is given by

$$\begin{aligned} s(t) &= 1.5s(t-1) - 0.7s(t-2) \\ &+ u(t-1) - 0.5u(t-2). \end{aligned} \quad (6-1)$$

Thus, we have $b_1 = 1.0$, $b_2 = -0.5$, $a_1 = -1.5$, and $a_2 = 0.7$. The input $\{u(t)\}$ is generated as

$$u(t) = 0.3u(t-1) + e(t) \quad (6-2)$$

where $\{e(t)\}$ is an i.i.d. binary sequence with values one and -1 , each with probability 0.5. Thus, we have $E\{e^2(t)\} = 1$, $\gamma_{3e} = 0$, and $\gamma_{4e} = -2$. The noisy input-output observations are given by (2-2) and (2-3). The mutually uncorrelated colored Gaussian noises $\{v_i(t)\}$ and $\{v_o(t)\}$ are generated as

$$v_i(t) = 0.8v_i(t-1) + f_i(t), \quad (6-3)$$

$$v_o(t) = -0.9v_o(t-1) + f_o(t) \quad (6-4)$$

where $\{f_i(t)\}$ and $\{f_o(t)\}$ are mutually independent i.i.d., zero-mean Gaussian sequences. The variances of $f_i(t)$ and

TABLE I

PARAMETER ESTIMATES: EXAMPLE 1, 30 MONTE CARLO RUNS, SNR = 20 dB, $N = 4000$ = INPUT-OUTPUT DATA RECORD LENGTH IN EACH RUN, σ = ONE STANDARD DEVIATION. [IT: INTEGRATED TRISPECTRUM BASED APPROACH, MSE: A MEAN-SQUARE ERROR CRITERION MINIMIZING THE PREDICTION ERRORS, SEE CRITERION (76) IN [39]]

Approach	Parameters	a_1	a_2	b_1	b_2
	True Values	-1.500	0.700	1.000	-0.500
	estimate statistics: N = 4000				
IT: Linear (Sec. 4.1)	mean	-1.4257	0.6364	0.9832	-0.4391
	σ	± 0.0242	± 0.0192	± 0.0136	± 0.0321
	RMS	(0.0782)	(0.0664)	(0.0173)	(0.0688)
IT: Nonlinear (Sec. 4.2)	mean	-1.4754	0.6774	0.9839	-0.4845
	σ	± 0.0075	± 0.0053	± 0.0083	± 0.0149
	RMS	(0.0257)	(0.0232)	(0.0182)	(0.0215)
[39]	mean	-1.3363	0.6189	0.9562	-0.4537
	σ	± 0.2366	± 0.1195	± 0.2093	± 0.2302
	RMS	(0.2877)	(0.2153)	(0.2138)	(0.2349)
[35]	mean	-1.5631	0.7494	1.0142	-0.5761
	σ	± 0.2585	± 0.2138	± 0.0113	± 0.2647
	RMS	(0.2661)	(0.2194)	(0.0182)	(0.2754)
MSE (from [39])	mean	-0.9490	0.2587	0.9902	0.0782
	σ	± 0.0076	± 0.0060	± 0.0108	± 0.0220
	RMS	(0.5511)	(0.4413)	(0.0145)	(0.5732)

$f_o(t)$ are selected to yield a signal to noise power ratio of 20 dB at the input sensor and the output sensor, respectively.

Thirty independent realizations of 4000 input-output data pairs were generated. The methods proposed in Section IV-A were applied using the integrated trispectrum, along with several existing approaches. For the linear estimator, we used $1 \leq l \leq L_B/2 - 1 = 30$, i.e., 30 frequency points in the interval $(0, \pi)$. The nonlinear estimator minimizing $J_N^{(IP)}(\theta)$ given by (4-12) was initialized by the linear estimator. The frequency points used in (4-12) were as that for the linear estimator.

Table I displays the arithmetic means, standard deviations and root mean-square (RMS) errors of the results of the various parameter estimators. The approach of [39] is based upon the fourth cumulant of a generalized error signal at zero lag, and the approach of [35] uses the fourth-order cross- and auto-cumulants of the input-output data at various lags. The results pertaining to the approach of [35] have been taken from [35]. For comparison the results of a least-squares criterion (labeled MSE in Table I) are also shown. It is seen that the MSE criterion relying only on the second-order statistics of the input-output record, produces a substantial bias (and low variance) for this example whereas the proposed frequency-domain approaches proposed in this paper yield very good results: the variance is comparable to the MSE criterion and the bias is much reduced. The proposed approaches perform much better than the approaches of [39] and [35].

B. Example 2 [34], [40]

The system to be identified is given by

$$\begin{aligned} s(t) &= 1.5s(t-1) - 0.7s(t-2) \\ &+ u(t) + 0.5u(t-1). \end{aligned} \quad (6-5)$$

Thus, we have $b_0 = 1.0$, $b_1 = 0.5$, $a_1 = -1.5$, and $a_2 = 0.7$.

The input $\{u(t)\}$ is obtained as

$$u(t) = e(t) - 0.2e(t-1) + 0.3e(t-2) \quad (6-6)$$

where, $\{e(t)\}$ is an i.i.d., zero-mean, unit variance (one-sided) exponential sequence. Thus, we have $E\{e^2(t)\} = 1$, and $\gamma_{3e} = 2$. The noisy input-output observations are given by (2-2) and (2-3). The mutually correlated colored Gaussian noises $\{v_i(t)\}$ and $\{v_o(t)\}$ are generated as

$$\begin{aligned} v_i(t) = & f_i(t) - 2.33f_i(t-1) + 0.75f_i(t-2) \\ & + 0.5f_i(t-3) + 0.3f_i(t-4) \\ & - 1.4f_i(t-5), \end{aligned} \quad (6-7)$$

$$\begin{aligned} v_o(t) = & v_i(t) + 0.2v_i(t-1) \\ & - 0.3v_i(t-2) + 0.4v_i(t-3) \end{aligned} \quad (6-8)$$

where $\{f_i(t)\}$ is an i.i.d., zero-mean Gaussian sequence. The variances of $v_i(t)$ and $v_o(t)$ are selected to yield a signal to noise variance ratio of 5 dB at the input sensor and the output sensor, respectively.

One hundred independent realizations of 4000 input-output data pairs were generated. The methods proposed in Sections III-B and IV-A (integrated bispectrum based) were applied along with several existing approaches. The nonlinear estimator minimizing $J_N^{(IP)}(\theta)$ given by (4-12) was initialized by the linear estimator of Section IV-A for which we used $1 \leq l \leq L_B/2 - 1 = 30$, i.e., 30 frequency points in the interval $(0, \pi)$. The frequency points used in (4-12) were identical to that for the linear estimator. A similar procedure was followed for minimization of $J_N^{(B)}(\theta)$ where we selected all the frequency points in the interior of D_{FFV} .

Table II displays the arithmetic means, standard deviations and root mean-square (RMS) errors of the results of the various parameter estimators. The approach of [31] uses the third-order cross- and auto-cumulants of the input-output data at various lags; so do the approaches of [34] and [40]. The results pertaining to the approaches of [34] and [40] have been quoted from [40]. It is seen from Table II that the frequency-domain approaches proposed in this paper yield the best results and the approaches of [31] and [34] the worst.

Note that $\hat{\theta}_N$ minimizing $J_N^{(B)}(\theta)$ is, asymptotically, a maximum likelihood parameter estimator under the restriction that only the third-order statistics are used. The other approaches ([31], [34], and [40]) also restrict themselves to third-order statistics of input-output record, but are not maximum likelihood. Since maximum likelihood parameter estimators are, in general, efficient, the proposed nonlinear estimator is expected to outperform other "nonoptimal" approaches. This seems to be the case for the simulation example considered here. Similar remarks are apply to integrated polyspectrum-based approaches.

VII. CONCLUSIONS

Two new classes of parametric frequency domain approaches were presented for estimation of the parameters of scalar, linear errors-in-variables models. One of the proposed classes of approaches consists of linear estimators where

TABLE II
PARAMETER ESTIMATES: EXAMPLE 1, 30 MONTE CARLO RUNS, SNR = 20 dB, $N = 4000$ = INPUT-OUTPUT DATA RECORD LENGTH IN EACH RUN, σ = ONE STANDARD DEVIATION, [IT]: INTEGRATED TRISPECTRUM BASED APPROACH, MSE: A MEAN-SQUARE ERROR CRITERION MINIMIZING THE PREDICTION ERRORS, SEE CRITERION (76) IN [39]

Approach	Parameters	a_1	a_2	b_0	b_1
	True Values	-1.500	0.700	1.000	0.500
	estimate statistics: N = 4000				
IB: Nonlinear (Sec. 4.2)	mean	-1.4815	0.6818	1.0333	0.4419
	σ	± 0.0047	± 0.0031	± 0.0192	± 0.0366
	RMS	(0.0191)	(0.0185)	(0.0384)	(0.0686)
Bisp.: Nonlinear (Sec. 3.2)	mean	-1.4555	0.6556	1.0647	0.4173
	σ	± 0.0245	± 0.0239	± 0.0435	± 0.0487
	RMS	(0.0508)	(0.0504)	(0.0779)	(0.0960)
[31]	mean	-1.8394	1.0351	1.0084	0.1581
	σ	± 0.4095	± 0.3834	± 0.0330	± 0.4232
	RMS	(0.5319)	(0.5092)	(0.0341)	(0.5453)
[34]	mean	-1.8200	0.8430	0.9700	0.3690
	σ	± 1.1680	± 1.4300	± 0.2450	± 0.9710
	RMS	(1.1740)	(1.4371)	(0.2627)	(0.9798)
[40] (Cross-cum.)	mean	-1.4444	0.6496	1.0596	0.5377
	σ	± 0.0472	± 0.0406	± 0.1278	± 0.1485
	RMS	(0.0729)	(0.0647)	(0.1410)	(0.1532)

the bispectrum or the integrated polyspectrum (bispectrum or trispectrum) of the input and the cross-bispectrum or the integrated cross-polyspectrum, respectively, of the input-output are exploited. Based on these polyspectra the system transfer function is first estimated at a number of frequencies exceeding one-half the number of unknown parameters. The estimated transfer function is then used to estimate the unknown parameters using an overdetermined linear system of equations. In the second class of approaches, quadratic transfer function matching criteria are optimized by using the results of the linear estimators as initial guesses. Both classes of the parameter estimators are shown to be consistent in any measurement noise that has symmetric probability density function when the bispectral approaches are used. The proposed parameter estimators are shown to be consistent in Gaussian measurement noise when trispectral approaches are used. The input to the system need not be a linear process but must have nonvanishing bispectrum or trispectrum. We emphasize that unlike the second-order statistics case, one can not, in general, model a stationary random process with a given higher-order cumulant spectrum as having been generated by driving a linear system with an i.i.d. sequence [27].

Computer simulation results were presented in support of the proposed approaches. Performance comparisons with several existing approaches show that the proposed approaches outperform them. On the theoretical side, the proposed estimators are consistent under sufficient conditions that are more general than that for existing approaches. Moreover the integrated polyspectrum-based approaches apply with almost equal computational and programming ease to both cases, those involving third-order statistics as well as those concerned with the fourth-order statistics. It should be noted, however, that higher-order statistics based methods typically yield high-variance estimates requiring "large" record sizes to reduce the variance.

We considered only SISO models. There are no essential difficulties in extending them to multiple-input multiple-output models although details remain to be worked out.

APPENDIX

In this Appendix we provide a proof of Theorem 2. First we need some auxiliary results.

Lemma 6. Under the Assumption Set I, it follows that

$$\begin{aligned} \lim_{N \rightarrow \infty} J_N^{(B)}(\theta) &= J_{\infty}^{(B)}(\theta) \\ &= \sum_{(m,n) \in \mathcal{S}_I \subset D_{xy}^0} [|H(e^{j(\omega_m + \omega_n)})| \theta_0] \\ &\quad - H(e^{j(\omega_m + \omega_n)} | \theta)|^2] / [\sigma_{mn}^2(\theta_0)] \quad (\text{A1}) \end{aligned}$$

1 p uniformly in θ for $\theta \in \Theta_c$ where $\Theta_c \subset \Theta$ is an arbitrary compact set, \mathcal{S}_I is a fixed collection of L bilfrequencies and $\sigma_{mn}^2(\theta_0) = q(m, n) S_r(\omega_m) S_r(\omega_n) S_y(\omega_m + \omega_n) / |B_{1,x}(\omega_m, \omega_n)|^2$.

Proof. By (3-17), (3-20), Lemma 1, and Slutsky's Theorem, we have

$$\begin{aligned} \lim_{N \rightarrow \infty} \sigma_{mn}^2 &= \sigma_{mn}^2(\theta_0) \quad 1 p, \quad (\text{A2}) \\ \lim_{N \rightarrow \infty} H(e^{j(\omega_m + \omega_n)}) \sigma_{mn}^{-1} &= H(e^{j(\omega_m + \omega_n)} | \theta_0) \\ \sigma_{mn}^{-1}(\theta_0) \quad 1 p \quad (\text{A3}) \end{aligned}$$

Consider

$$\begin{aligned} D_{mn}(\theta) &= |H(e^{j(\omega_m + \omega_n)}) - H(e^{j(\omega_m + \omega_n)} | \theta)|^2 \sigma_{mn}^{-2} \\ &\quad - |H(e^{j(\omega_m + \omega_n)} | \theta_0) - H(e^{j(\omega_m + \omega_n)} | \theta)|^2 \sigma_{mn}^{-2} \quad (\text{A4}) \end{aligned}$$

$$\begin{aligned} |H(e^{j(\omega_m + \omega_n)} | \theta)|^2 \sigma_{mn}^{-2} &= |H(e^{j(\omega_m + \omega_n)} | \theta_0)|^2 \sigma_{mn}^{-2} \\ &\quad + 2 \operatorname{Re} \{ H(e^{j(\omega_m + \omega_n)} | \theta) [H(e^{j(\omega_m + \omega_n)} | \theta_0)]^* \sigma_{mn}^{-2} \} \quad (\text{A5}) \end{aligned}$$

By compactness of Θ_c and continuity of $H(e^{j(\omega_m + \omega_n)} | \theta)$ in θ as well as in ω_m and ω_n , we have

$$\sup_{\omega_m \in [-\pi, \pi]} \sup_{\omega_n \in [-\pi, \pi]} \sup_{\theta \in \Theta_c} |H(e^{j(\omega_m + \omega_n)} | \theta)| \leq M < \infty \quad (\text{A6})$$

By (A2), (A3), and Slutsky's Theorem, given any ϵ and $\delta > 0$ there exists an integer $N(\epsilon, \delta)$ such that

$$\begin{aligned} P\{ ||[H(e^{j(\omega_m + \omega_n)})]^2 - |H(e^{j(\omega_m + \omega_n)} | \theta_0)|^2] \sigma_{mn}^{-2} | < \epsilon \} \\ > 1 \quad (\text{A7}) \end{aligned}$$

and

$$\begin{aligned} P\{ ||[H(e^{j(\omega_m + \omega_n)}) - H(e^{j(\omega_m + \omega_n)} | \theta_0)] \sigma_{mn}^{-2} | < \epsilon \} \\ > 1 - \delta \quad (\text{A8}) \end{aligned}$$

for every $N \geq N(\epsilon, \delta)$. Hence it follows that given any $\epsilon_1 (= \epsilon + 2M\epsilon)$ and $\delta > 0$ there exists an integer $N(\epsilon_1, \delta)$ such that

$$P\{|D_{mn}(\theta)| < \epsilon_1\} > 1 - \delta \quad \forall \theta \in \Theta_c, \quad \forall N \geq N(\epsilon_1, \delta) \quad (\text{A9})$$

That is, convergence of $D_{mn}(\theta)$ is uniform in θ . The desired result now follows by using the uniform convergence of $D_{mn}(\theta)$, Slutsky's Theorem, and noting that

$$\begin{aligned} |J_{\infty}^{(B)}(\theta) - J_N^{(B)}(\theta)| &\leq \sum_{(m,n) \in \mathcal{S}_I \subset D_{xy}^0} |D_{mn}(\theta)| \leq L\epsilon_1 \\ & \quad (\text{A10}) \end{aligned}$$

if $|D_{mn}(\theta)| \leq \epsilon_1$ \square

Lemma 7. Under the Assumption Set I, we have

$$J_{\infty}^{(B)}(\theta) = 0 \quad \text{for any } \theta \in \Theta \Leftrightarrow \theta \in D_0$$

Proof. From (A1)

$$\begin{aligned} J_{\infty}^{(B)}(\theta) = 0 &\Rightarrow H(e^{j(\omega_m + \omega_n)} | \theta_0) = H(e^{j(\omega_m + \omega_n)} | \theta) \\ &\quad \text{for } (m, n) \in \mathcal{S}_I \subset D_{xy}^0 \\ &\Rightarrow H(e^{j\omega} | \theta_0) = H(e^{j\omega} | \theta) \quad \forall \omega \in (-\pi, \pi], \text{ by Lemma 5} \\ &\Rightarrow \theta \in D_0 \end{aligned}$$

Conversely

$$\begin{aligned} \theta \in D_0 &\Rightarrow H(e^{j(\omega_m + \omega_n)} | \theta_0) = H(e^{j(\omega_m + \omega_n)} | \theta) \\ &\quad \text{for } (m, n) \in \mathcal{S}_I \subset D_{xy}^0 \\ &\Rightarrow J_{\infty}^{(B)}(\theta) = 0 \end{aligned}$$

This completes the proof \square

Proof of Theorem 2. Define a set

$$\Theta(\rho) = \Theta_c - S(D_0, \rho)$$

where

$$S(D_0, \rho) = \{\theta | \theta \in \Theta, \inf_{\bar{\theta} \in D_0} \|\theta - \bar{\theta}\| < \rho\}$$

Since $S(D_0, \rho)$ is open, $\Theta(\rho)$ is a closed subset of a compact set $\Theta_c \subset \Theta$, hence compact. By continuity of $J_{\infty}^{(B)}(\theta)$ in θ and compactness of $\Theta(\rho)$, there exists some $\theta^* \in \Theta(\rho)$ such that

$$\begin{aligned} \inf_{\theta \in \Theta(\rho)} J_{\infty}^{(B)}(\theta) &= J_{\infty}^{(B)}(\theta^*) = \mu(\rho) > 0 \\ &= J_{\infty}^{(B)}(\theta_0) \quad (\text{A11}) \end{aligned}$$

where we have used Lemma 7 in deducing that $\mu(\rho) > 0$. By Lemma 6, given any $\epsilon \leq \mu(\rho)/3$ and $\delta > 0$ there exists an integer $N_1(\epsilon, \delta)$ such that $\forall \theta \in \Theta(\rho)$ and $\forall N \geq N_1(\epsilon, \delta)$

$$P\{J_N^{(B)}(\theta) > J_{\infty}^{(B)}(\theta) - \epsilon \geq \mu(\rho) - \epsilon\} > 1 - \delta \quad (\text{A12})$$

The above equation may be rewritten as

$$\begin{aligned} P\{ \inf_{\theta \in \Theta(\rho)} J_N^{(B)}(\theta) > \mu(\rho) - \epsilon \} &> 1 - \delta \\ \forall N &\geq N_1(\epsilon, \delta) \quad (\text{A13}) \end{aligned}$$

Similarly, by Lemma 6, given any $\epsilon \leq \mu(\rho)/3$ and $\delta > 0$ there exists an integer $N_2(\epsilon, \delta)$ such that $\forall \theta \in \Theta_c$ and $\forall N \geq N_2(\epsilon, \delta)$ we have

$$P\{J_N^{(B)}(\theta) < J_{\infty}^{(B)}(\theta) + \epsilon\} > 1 - \delta \quad (\text{A14})$$

In particular, we have

$$\begin{aligned} P\{J_N^{(B)}(\theta_0) < J_{\infty}^{(B)}(\theta_0) + \epsilon = \epsilon\} &> 1 - \delta \\ \forall N &\geq N_2(\epsilon, \delta) \quad (\text{A15}) \end{aligned}$$

Since

$$\inf_{\theta \in \Theta_c} J_N^{(B)}(\theta) \leq J_N^{(B)}(\theta_0)$$

however, it follows from (A14) and (A15) that

$$\begin{aligned} P\left\{\inf_{\theta \in \Theta_c} J_N^{(B)}(\theta) < \epsilon \leq \mu(\rho)/3\right\} &> 1 - \delta \\ \forall N \geq N_2(\epsilon, \delta). \end{aligned} \quad (\text{A16})$$

Consider $\hat{\theta}_N$ as defined in (5-13). Then we have

$$\begin{aligned} P\{\hat{\theta}_N \in S(D_0, \rho)\} &= P\left\{\inf_{\theta \in \Theta_c} J_N^{(B)}(\theta) < \inf_{\theta \in \Theta(\rho)} J_N^{(B)}(\theta)\right\} \\ &\geq P\left\{\inf_{\theta \in \Theta_c} J_N^{(B)}(\theta) < \epsilon \leq \mu(\rho)/3, \right. \\ &\quad \left. \mu(\rho) - \epsilon < \inf_{\theta \in \Theta(\rho)} J_N^{(B)}(\theta)\right\} \\ &\geq 1 - P\left\{\inf_{\theta \in \Theta_c} J_N^{(B)}(\theta) \geq \mu(\rho)/3 \geq \epsilon\right\} \\ &\quad - P\left\{\inf_{\theta \in \Theta(\rho)} J_N^{(B)}(\theta) \leq \mu(\rho) - \epsilon\right\} \\ &\geq 1 - 2\delta \quad \forall N \geq N(\epsilon, \delta) = \max\{N_1(\epsilon, \delta), N_2(\epsilon, \delta)\} \end{aligned} \quad (\text{A17})$$

where we used (A13) and (A16) to obtain (A17). This proves the desired result. \square

REFERENCES

- [1] T. Söderström, "Spectral decomposition with application to identification," in *Numerical Techniques for Stochastic Systems*, F. Archetti and M. Cugiani, Eds. Amsterdam: North-Holland, 1980.
- [2] ———, "Identification of stochastic linear systems in presence of input noise," *Automatica*, vol. 17, pp. 713–725, Sep. 1981.
- [3] B. D. O. Anderson and M. Deistler, "Identifiability in dynamic errors-in-variables models," *J. Time Series Analysis*, vol. 5, pp. 1–13, 1984.
- [4] B. D. O. Anderson, "Identification of scalar errors-in-variables models with dynamics," *Automatica*, vol. 21, pp. 709–716, 1985.
- [5] M. Deistler, "Linear errors-in-variables models," in *Time Series and Linear Systems* (Lecture Notes in Control and Information Sciences) S. Bittanti, Ed. vol. 86. Berlin: Springer-Verlag, 1986, pp. 37–86.
- [6] M. Deistler and B. D. O. Anderson, "Linear dynamic errors-in-variables models: Some structure theory," *J. Econometrics*, vol. 41, pp. 39–63, 1989.
- [7] T. Söderström and P. Stoica, *System Identification*. London: Prentice-Hall Int., 1989.
- [8] P. Stoica and A. Nehorai, "On the uniqueness of prediction error models for systems with noisy input-output data," *Automatica*, vol. 23, pp. 541–543, 1987.
- [9] M. Green and B. D. O. Anderson, "Identification of multivariable errors-in-variables models with dynamics," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 467–471, 1986.
- [10] A. G. Piersol, "Time delay estimation using phase data," *IEEE Trans. Acoustics, Speech, Sig. Proces.*, vol. ASSP-29, pp. 471–477, June 1981.
- [11] W. S. Burdick, *Underwater Acoustic System Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
- [12] C. L. Nikias and R. Pan, "Time delay estimation in unknown Gaussian spatially correlated noise," *IEEE Trans. Acoustics, Speech, Sig. Proces.*, vol. 36, pp. 1706–1714, Nov. 1988.
- [13] J. K. Tugnait, "Time delay estimation in unknown spatially correlated Gaussian noise using higher-order statistics," in *Proc. 23rd Asilomar Conf. Sig., Syst., Computers*, Pacific Grove, CA, Oct. 30–Nov. 1, 1989, pp. 211–215.
- [14] H. Akaike, "On the use of non-Gaussian process in the identification of a linear dynamic system," *Ann. Inst. Statistical Math.*, vol. 18, pp. 269–276, 1966.
- [15] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
- [16] M. Deistler, "Linear dynamic errors-in-variables models," in *Essays In Time Series And Allied Processes*, J. Applied Probability, J. Gani and M. B. Priestley, Ed., vol. 23A. 1986, pp. 23–40.
- [17] V. Solo, "Identifiability of time series models with errors in variables," in *Essays In Time Series And Allied Processes*, J. Applied Probability, J. Gani and M. B. Priestley, Ed., vol. 23A. 1986, pp. 63–74.
- [18] R. E. Kalman, "Identifiability and modeling in econometrics," in *Developments in Statistics*, P. R. Krishnaiah, Ed. vol. 4. New York: Academic, 1983, pp. 97–136.
- [19] K. J. Åström and P. Eykhoff, "System identification—A survey," *Automatica*, vol. 7, pp. 123–162, 1971.
- [20] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
- [21] T. Subba Rao and M. M. Gabr, *An Introduction to Bispectral Analysis and Bilinear Time Series Models*. New York: Springer-Verlag, 1984.
- [22] K. S. Lii and M. Rosenblatt, "Deconvolution and estimation of transfer function phase and coefficients for nongaussian linear processes," *Ann. Statistics*, vol. 10, pp. 1195–1208, 1982.
- [23] D. R. Brillinger, "An introduction to polyspectra," *Annals Math. Statistics*, vol. 36, pp. 1351–1374, 1965.
- [24] M. Rosenblatt, *Stationary Sequences and Random Fields*. Boston: Birkhäuser, 1985.
- [25] M. J. Hinich, "Testing for Gaussianity and linearity of a stationary time series," *J. Time Series Analysis*, vol. 3, no. 3, pp. 169–176, 1982.
- [26] B. Y. Hamon and E. J. Hannan, "Spectral estimation of time delay for dispersive and nondispersive systems," *J. Royal Statist. Soc. Ser. C (Applied Statistics)*, vol. 23, pp. 134–142, 1974.
- [27] A. M. Tekalp and A. T. Erdem, "Higher-order spectrum factorization in one and two dimensions with applications in signal modeling and non-minimum phase system identification," *IEEE Trans. Acoustics, Speech, Sig. Proces.*, vol. 37, pp. 1537–1549, Oct. 1989.
- [28] J. K. Tugnait, "Stochastic system identification with noisy input using cumulant statistics," in *Proc. 29th IEEE Conf. Decis. Contr.*, Honolulu, HI, Dec. 5–7, 1990, pp. 1080–1085.
- [29] M. Deistler, "Symmetric modeling in system identification," in *Three Decades of Mathematical System Theory* (Lecture Notes in Control & Information Sciences), H. Nijmeijer and J. M. Schumacher, Eds. Berlin: Springer-Verlag, 1989.
- [30] K. S. Lii and K. N. Helland, "Cross-bispectrum computation and variance estimation," *ACM Trans. Math. Software*, vol. 7, pp. 284–294, Sep. 1981.
- [31] Y. Inouye and H. Tsuchiya, "Identification of linear systems using input-output cumulants," *Int. J. Contr.*, vol. 53, pp. 1431–1448, 1991.
- [32] J. A. Cadzow and O. M. Solomon, Jr., "Algebraic approach to system identification," *IEEE Trans. Acoustics, Speech, Sig. Proces.*, vol. ASSP-34, pp. 462–469, June 1986.
- [33] J. K. Tugnait and Y. Ye, "Stochastic system identification with noisy input-output measurements," in *Proc. 26th Annual Asilomar Conf. Sig. Syst. Computers*, Pacific Grove, CA, Oct. 26–28, 1992, pp. 741–745.
- [34] A. Delopoulos and G. B. Giannakis, "Strongly consistent output only and input/output identification in the presence of Gaussian noise," in *Proc. 1991 ICASSP*, Toronto, Canada, May 14–17, 1991, pp. 3521–3524.
- [35] Y. Inouye and Y. Suga, "Identification of linear systems with noisy input using input-output cumulants," *Int. J. Contr.*, vol. 59, pp. 1231–1253, May 1994.
- [36] K. L. Chung, *A Course In Probability Theory*. New York: Harcourt, Brace, World Inc., 1968.
- [37] E. J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*. New York: Wiley, 1988.
- [38] G. J. Dobeck, V. K. Jain, K. W. Watkinson, and D. E. Humphreys, "Identification of submerged vehicle dynamics through a generalized least squares method," in *Proc. 1976 IEEE Conf. Decis. Contr.*, Clearwater, FL, Dec. 1976, pp. 78–83.
- [39] J. K. Tugnait, "Stochastic system identification with noisy input using cumulant statistics," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 476–485, Apr. 1992.
- [40] J. M. M. Anderson and G. B. Giannakis, "Noisy input/output system identification using cumulants and the Steiglitz-McBride algorithm," in *Proc. 25th Asilomar Conf. Sig., Syst., Computers*, Pacific Grove, CA, Nov. 1991, pp. 608–612.
- [41] M. J. Hinich and G. R. Wilson, "Time delay estimation using the cross bispectrum," *IEEE Trans. Sig. Proces.*, vol. 40, pp. 106–113, Jan. 1992.
- [42] D. R. Brillinger, *Time Series Data Analysis and Theory*. New York: Holt, Rinehart and Winston, 1975.
- [43] D. R. Brillinger and M. Rosenblatt, "Asymptotic theory of estimates of kth order spectra," in *Spectral Analysis of Time Series*, B. Harris, ed. New York: Wiley, 1967, pp. 153–188.
- [44] J. K. Tugnait, "On time delay estimation with unknown spatially correlated Gaussian noise using fourth order cumulants and cross cumulants," *IEEE Trans. Sig. Proces.*, vol. 39, pp. 1258–1267, June 1991.
- [45] G. C. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1984.



Jitendra K. Tugnait (M'79-SM'93-F'94) was born in Jabalpur, India on Dec. 3, 1950. He received the B.Sc. (Hons.) degree in electronics and electrical communication engineering from the Punjab Engineering College, Chandigarh, India in 1971, the M.S. and the E.E. degrees from Syracuse University, Syracuse, NY and the Ph.D. degree from the University of Illinois, Urbana-Champaign in 1973, 1974, and 1978, respectively, all in electrical engineering.

From 1978 to 1982 he was an Assistant Professor of Electrical and Computer Engineering at the University of Iowa, Iowa City. He has also been associated with the School of Radar Studies, Indian Institute of Technology, New Delhi, India, and Space Applications Centre, Ahmedabad, India, during 1975-1976. He was with the Long Range Research Division of the Exxon Production Research Company, Houston, TX, from June 1982 to Sept. 1989. He joined the Department of Electrical Engineering, Auburn University, Auburn, AL, in 1989 as a Professor. His research interests are in statistical signal processing and stochastic systems analysis (with emphasis on higher-order statistics), with applications to communications, control and image/signal processing.

Dr. Tugnait served as an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL during 1985 and 1986, as an appointed member of the Board of Governors of the IEEE Control Systems Society in 1986, and as a member of the Operating Committee of the 1991 CDC in Brighton, UK. He is currently a member of the statistical signal and array processing technical committee of the IEEE Signal Processing Society and is an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING.



Yisong Ye (S'90-M'93) was born in Shanghai, People's Republic of China on April 30, 1965. He received the B.E.E. degree from Shanghai Jiao Tong University, Shanghai, China in 1986 and the M.S.E.E. degree from Auburn University, Auburn, AL in 1992. He is currently a Ph.D. candidate in electrical engineering at Auburn University.

Mr. Ye's research interests include system identification, signal processing and communications.

Minimum Bias Priors for Estimating Parameters of Additive Terms in State-Space Models

Bertrand Hochwald, *Student Member, IEEE*, and Arye Nehorai, *Fellow, IEEE*

Abstract—We treat the problem of estimating parameters of additive terms, sometimes called bias terms, in state-space models. We consider models that depend linearly on the state but possibly nonlinearly on the parameters, where both the state and observation are corrupted by additive noise. A prior density for the parameters is introduced that, when combined with the likelihood function to form a posterior density, minimizes the bias of the posterior mean. The result is a useful prior based on ignorance. Two examples and simulations illustrate the use of the prior.

I. INTRODUCTION

THE problem of determining unknown additive terms in state-space models has been addressed by many authors; for some early examples, see [4] and [8]. These models appear frequently when, for example, noise of unknown mean is encountered, sensors with systematic errors are used, or moving objects are being tracked. We consider the problem of obtaining the posterior mean estimate of parameters that appear nonlinearly in the additive terms. The posterior mean is important because of its wide use in estimation theory and its close connection to the Kalman filter estimate of the additive terms [4]. For example, when the parameters appear linearly in the additive terms and have a jointly Gaussian prior distribution, the Kalman and posterior mean estimates coincide.

We then suggest a method to choose a prior density for the parameters that conveys ignorance. When information about parameters is unavailable, there are various criteria for choosing noninformative priors. One is convenience: If we assume that the parameters have a distribution amenable to simple recursive updating equations, then estimates are often easy to compute. Another is lack of favoritism: If we assume the parameters have a uniform distribution over some interval, then it is widely accepted that we are not giving preference to any particular parameter value. There are also information theoretic criteria that, for example, maximize the Kullback–Liebler–Lindley distance between the posterior density and the prior [6], [9]. These criteria, however, while broad in scope, often cannot predict precisely what

effect the priors they recommend will have on a particular estimate.

By focusing on the posterior mean, we are able to show the effect of any prior density, and we choose a prior to minimize bias (to be defined shortly). If the true prior is used to construct a posterior density, then the posterior mean has the well-known minimum mean-square error property. This optimality is measured as an average over possible realizations of the parameters drawn in accordance with their distribution. If, however, a prior other than the true one is used to construct a posterior density, then this ensemble average property of the posterior mean is generally lost.

In this paper we obtain a prior density function that satisfies an optimality condition within a realization instead: When used to make a posterior density function of the parameters, the prior, asymptotically in the number of observations, minimizes the posterior mean's bias over all possible priors. Let θ be the parameter we wish to estimate and $\hat{\theta}_t$ be the posterior mean computed at time t using the prior f_θ . Supposing θ_0 denotes the true value of θ and E_{θ_0} denotes expectation conditioned on θ_0 , we show that $E_{\theta_0}\hat{\theta}_t = \theta_0 + O(t^{-1})$, as $t \rightarrow \infty$, for most f_θ . But the proposed f_θ (which can depend on t) satisfies $E_{\theta_0}\hat{\theta}_t = \theta_0 + o(t^{-1})$. Hence the name “minimum bias prior.”

As more observations become available, the density we propose varies. Contrast this with the traditional prior density that reflects what is known about the parameters before observations are made. We draw on the work of Hartigan [5], in which minimum bias priors are first developed. He determines that (asymptotically) the first-order effect of the prior appears in the posterior mean. Effects on the higher order moments of the posterior density are negligible by comparison. The minimum bias prior makes the first-order effect as small as possible.

In Section II the state-space model with parameterized additive terms is described and the estimation problem formulated. We show how to recursively calculate the unnormalized posterior density function of the parameters in a simple and efficient manner. In Section III we introduce the minimum bias prior. The prior is developed in [5] under the assumption that identically distributed observations of the parameters are available. Our model will not satisfy this assumption, requiring us to extend the prior's theory. A method is also given to approximate the prior with an adaptive scheme, since an exact closed-form solution cannot always be obtained. The prior used in the adaptive scheme is, however, particularly

Manuscript received September 24, 1993; revised June 10, 1994. Recommended by Past Associate-Editor, B. Pasik-Duncan. This work was supported by Air Force Office of Scientific Research Grant AFOSR-90-0164, Office of Naval Research Grant N00014-91-J-1298, and National Science Foundation Grant MIP-9122753.

The authors are with the Department of Electrical Engineering, Yale University, New Haven, CT 06520 USA.

IEEE Log Number 9409417.

simple. Two examples and simulations demonstrate the results in Section IV, and we conclude in Section V.

II. MODEL AND PROBLEM FORMULATION

Consider the state-space model

$$x_t = Ax_{t-1} + Bg(\theta) + w_{t-1}, \quad (2.1a)$$

$$y_t = Cx_t + Dh(\theta) + v_t \quad (2.1b)$$

where $t = 1, 2, \dots$ is the time index. The vector $y_t \in \mathbb{R}^{m \times 1}$ is the observation, $x_t \in \mathbb{R}^{n \times 1}$ is the state, and $\theta \triangleq [\theta_1, \dots, \theta_p]^T \in \mathbb{R}^{p \times 1}$ is the unknown parameter vector. The noise vectors $w_t \sim \mathcal{N}(0, Q)$, $v_t \sim \mathcal{N}(0, R)$ (have zero-mean Gaussian distributions with respective variances) and w_s are assumed to be independent of v_t for all s and t . It is supposed that $Q \in \mathbb{R}^{n \times n}$ is positive semidefinite and $R \in \mathbb{R}^{m \times m}$ is positive definite for all t . The parameter vector θ and initial state x_0 are random and independent. Furthermore, we suppose that $g(\theta) \in \mathbb{R}^{q \times 1}$ and $h(\theta) \in \mathbb{R}^{r \times 1}$ are known functions of θ representing the additive terms. All matrices are assumed to have the appropriate dimensions and to be known.

A common approach to simultaneous estimation of the state and parameter vectors is to extend the state-vector to include the additive terms by letting $\tilde{x}_t \triangleq [x_t^T, g^T(\theta), h^T(\theta)]^T$, whence

$$\begin{bmatrix} A & B & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \tilde{x}_{t-1} + \begin{bmatrix} w_{t-1} \\ 0 \\ 0 \end{bmatrix} \quad (2.2a)$$

$$y_t = [C \quad 0 \quad D] \tilde{x}_t + v_t. \quad (2.2b)$$

If we assume that the initial vector \tilde{x}_0 has some covariance matrix, then the model is in a form suitable for applying the Kalman filtering equations to estimate \tilde{x}_t [4]. This provides recursive estimates of g and h . If $\theta \mapsto (g(\theta), h(\theta))$ is one-to-one we may invert the estimates of g and h to get estimates of θ .

While this method is attractive because of its computational simplicity, we may not want to be forced into assigning g and h a covariance that can, at best, be a guess. This is true especially because the parameter of interest is θ and not $(g(\theta), h(\theta))$. Moreover, the criterion we consider in this paper is bias, and the extended-state Kalman filter, which essentially obtains the posterior mean of g and h under the assumption that they have a joint-Gaussian prior, may have a bias when estimating θ that is higher than that obtainable with the minimum bias prior. (Due to the nonlinearities in g and h , this holds even when g and h can be estimated with no bias.) Example 2 in Section IV demonstrates this.

We use the term realization to denote a series of measurements obtained via (2.1) in which θ is fixed at θ_0 . When we speak of an average within a realization, the expectation is over all possible values of the observations (assuming θ is fixed) and the relevant density is $f_{Y^t|\theta_0}$, where Y^t is shorthand for y_1, \dots, y_t . This expectation is denoted E_{θ_0} .

Suppose a collection of t observations from one realization is available and $\hat{\theta}_t = \int \theta f_{\theta|Y^t} d\theta$. Our goal is to minimize

the asymptotic bias, $E_{\theta_0} \hat{\theta}_t - \theta_0$, by modifying $f_{\theta|Y^t}$, through selecting an appropriate prior, f_θ . We will use f_θ to denote a prior density for the parameters, whether it is the true prior or not. The true prior plays no role in the bias criterion so there should be no source of confusion, and it will not matter what law is used to choose θ . Similarly, we will not distinguish between posterior densities and means computed using different f_θ .

Assumptions:

A1) $f_{x_0|\theta} = \mathcal{N}(m_0, P_0)$.

A2) The noise is "white," implying that $f_{x_t|x_{t-1}, Y^{t-1}, \theta} = f_{x_t|x_{t-1}, \theta}$ and $f_{y_t|x_t, Y^{t-1}, \theta} = f_{y_t|x_t, \theta}$.

We begin by computing the joint posterior density function of x_t and θ . Define $l(\theta) = [g^T(\theta), h^T(\theta)]^T$.

Lemma 1:

$$f_{x_t, \theta|Y^t} = f_{x_t|Y^t, \theta} \cdot f_{\theta|Y^t} \quad (2.3)$$

where

$$f_{x_t|Y^t, \theta} = \mathcal{N}(M_t l(\theta) + z_t, P_t), \quad (2.4)$$

$$f_{\theta|Y^t} \propto f_\theta \cdot \exp \{ \alpha_t^T l(\theta) + l^T(\theta) \Lambda_t l(\theta) \} \quad (2.5)$$

and z_t, α_t, M_t, P_t , and Λ_t may be computed recursively.

Proof: See Appendix A.

Given t observations, the posterior mean of x_t and θ is $\int [x_t^T, \theta^T]^T f_{x_t, \theta|Y^t} d x_t d \theta$; (2.3)–(2.5) yield

$$\hat{x}_t = \frac{1}{c_t} M_t \int l(\theta) f_\theta \cdot \exp \{ \alpha_t^T l(\theta) + l^T(\theta) \Lambda_t l(\theta) \} \cdot d\theta + z_t, \quad (2.6a)$$

$$\hat{\theta}_t = \frac{1}{c_t} \int \theta f_\theta \cdot \exp \{ \alpha_t^T l(\theta) + l^T(\theta) \Lambda_t l(\theta) \} d\theta \quad (2.6b)$$

where $c_t = \int f_\theta \cdot \exp \{ \alpha_t^T l(\theta) + l^T(\theta) \Lambda_t l(\theta) \} d\theta$ is a normalizing constant. Observe the notation: \hat{x}_t the estimate of x_t , and $\hat{\theta}_t$ the estimate of the constant θ_0 .

In what follows we concern ourselves exclusively with $\hat{\theta}_t$ and examining its behavior as a functional of f_θ . Generally, $\hat{\theta}_t$ converges in probability to θ_0 as $t \rightarrow \infty$. We will, in fact, give conditions that imply convergence of θ_t to θ_0 in expectation, and choose f_θ to optimize the speed with which this happens.

III. MINIMUM BIAS PRIORS

The minimum bias prior minimizes the first-order effect of the prior on the posterior density [5]. When θ is a scalar we show that there is a closed-form solution for the prior. When θ is a vector, we can show only that it satisfies a certain set of differential equations that do not necessarily yield a closed-form solution. To handle this, we propose a simple adaptive approximation of the prior.

The minimum bias prior depends on the likelihood function. Jeffreys [7] first suggested using the likelihood function to design priors, with the Jeffreys prior being a well-known example. It turns out that in one dimension the minimum bias prior is the square of the Jeffreys prior (see [2] for other uses of the Jeffreys prior). In more than one dimension, the

minimum bias prior and the square of the Jeffreys prior no longer necessarily coincide.

We would like to apply the results of Hartigan in [5] but some justification is needed. He considers a model for which independent identically distributed observations are available. That is, in his case, $\log f_{Y^t|\theta} = \sum_{r=1}^t \log f_{Y_r|\theta}$ and $\{\log f_{Y_r|\theta}\}_{r=1,2,\dots}$ is a collection of independent identically distributed random variables; many of his results are based on the asymptotic behavior of this sum. In our case, $\log f_{Y^t|\theta} = \sum_{r=1}^t \log f_{Y_r|Y^{r-1},\theta}$, and the summands are not identically distributed. As we shall momentarily see, however, the summands are independent and have the same distribution asymptotically, allowing us to use the results of [5] anyway.

The first step is to observe that from (A.2), (A.6), and (A.7), $f_{Y_t|Y^{t-1},\theta} = \mathcal{N}(N_t l(\theta) + C A z_{t-1}, S_t)$, where $N_t = C A M_{t-1} + [C B, D]$ and $S_t = C(A P_{t-1} A^T + Q) C^T + R$. The asymptotic behavior of $\log f_{Y_t|Y^{t-1},\theta}$ is principally governed by N_t and S_t which are shown to converge under the following assumption.

Assumption

A3) The pair (C, A) is detectable and $(A, Q^{1/2})$ is stabilizable [1].

Lemma 2: The matrices N_t and S_t tend to limits exponentially fast as $t \rightarrow \infty$.

Proof We have $S_t = A P_{t|t-1} A^T + R$ where $P_{t|t-1} = A P_{t-1} A^T + Q$. That $P_{t|t-1}$ tends to a unique positive definite matrix exponentially fast under Assumption A3) can be found in [1]. Therefore, the lemma is proved for S_t . To prove the result for N_t , we observe that it is sufficient to prove the result for M_t . From (A.4), $M_t = (A - K_t C A) M_{t-1} + [(I - K_t C) B, -K_t D]$, where $K_t = P_{t|t-1} C^T S_t^{-1}$. It is well known that K_t converges exponentially quickly to some K_∞ such that the eigenvalues of $A - A K_\infty C$ have modulus less than one, whence $A - K_t C A$ has eigenvalues with modulus less than one for t sufficiently large. Therefore M_t converges at an exponential rate as well. \square

Define

$$\eta_t(\theta) = \log f_{Y_t|Y^{t-1},\theta}, \quad t = 1, 2, \dots, \\ = -\frac{1}{2} [\log \det 2\pi S_t + (y_t - N_t l(\theta) - C A z_{t-1})^T \\ \cdot S_t^{-1} (y_t - N_t l(\theta) - C A z_{t-1})] \quad (3.1)$$

$N_\infty \triangleq \lim_{t \rightarrow \infty} N_t$, $S_\infty \triangleq \lim_{t \rightarrow \infty} S_t$, and $\eta_\infty(\theta) \triangleq -\frac{1}{2} [\log \det 2\pi S_\infty + \varepsilon_\infty^T(\theta) S_\infty^{-1} \varepsilon_\infty(\theta)]$ where $\varepsilon_\infty \sim \mathcal{N}(N_\infty [l(\theta_0) - l(\theta)], S_\infty)$. We use Lemma 2 to show that $\eta_t(\theta)$ converges in distribution to $\eta_\infty(\theta)$ exponentially quickly.

Lemma 3

- i) The random variables η_s and η_t are independent for $s \neq t$.
- ii) The random variables η_t converge in distribution, as $t \rightarrow \infty$, to η_∞ . Furthermore, the expectation of any power of any derivative of η_t with respect to θ tends exponentially quickly to the expectation of the same power and derivative of η_∞ .

Proof: See Appendix B.

Lemma 3 shows us that $\log f_{Y_t|Y^{t-1},\theta}$, $t = 1, 2, \dots$, are

almost identically distributed. The fact that they are not exactly identically distributed turns out to be of no consequence. We show this in the lemma that follows, extending Lemma 1 of Hartigan's paper to our model. First, some regularity conditions on f_θ and $\eta_t(\theta)$ similar to those appearing in [5] are stated.

Assumptions

- B1) θ_0 is an interior point of Θ where $\Theta \subset \mathbb{R}^p$ is closed.
- B2) $l(\theta)$ is continuous for $\theta \in \Theta$ and continuously four times differentiable in some neighborhood of θ_0 .
- B3) The parameters are identifiable: For $\theta \neq \theta_0$, $|N_t[l(\theta) - l(\theta_0)]|$ is bounded from below by a positive constant, uniformly in t .
- B4) $\log f_\theta$, $(\partial/\partial\theta) \log f_\theta$ and $(\partial^2/\partial\theta\partial\theta^T) \log f_\theta$ exist and are continuous in a neighborhood of θ_0 .
- B5) The prior density f_θ may be improper [$\notin L^1(\Theta)$], but $\int_\Theta f_{Y_t|Y^{t-1},\theta} f_\theta d\theta$ exists and is a bounded function of y_t , uniformly in Y^{t-1} and t .
- B6) Either Θ is bounded, or for some $k > 0$, $\sup_{|\theta| > k} [\eta_t(\theta) - \eta_t(\theta_0)]$ has first moment bounded from above by a negative constant and bounded fourth moment, both uniformly in t .

Lemma 4 For any $c > 0$ there exists a $\delta > 0$ such that $f_{Y^t|\theta}/f_{Y^t|\theta_0}$ is $O(e^{-t\delta})$ with probability $1 - O(t^{-2})$, uniformly in θ for $|\theta - \theta_0| > c$.

Proof See Appendix C.

To extend his results completely we need to also show that

$$\sum_{i=1}^t \frac{\partial^s \eta_i(\theta)}{\partial \theta^s} \bigg|_{\theta=\theta_0} = t c_{i,s} + O_p(t^1) \\ i = 1, \dots, p, s = 1, \dots, 4$$

for some $c_{i,s} \in \mathbb{R}$. But that this is true with $c_{i,s} = E_{\theta_0} \{\partial^s \eta_\infty(\theta) / \partial \theta^s |_{\theta=\theta_0}\}$ follows directly from Chebyshev's inequality and Lemma 3. We are now in a position to use the results on minimum bias priors.

Let $V(d, \theta)$ be a smooth nonnegative loss function with unique minimum at $\theta = d$ and

$$\theta_t = \arg \min_d \int V(d, \theta) f_{\theta|Y^t} d\theta.$$

For a scalar θ , under assumptions B1–B6 we have the asymptotic expansion

$$\arg \min_{\theta \in \Theta} E_{\theta_0} V(\hat{\theta}_t, \theta) = \theta_0 - p_2^{-1} [q_1 - (p_{12} + p_1)/p_2 \\ - \frac{1}{2} (V_{12} + V_{21})/V_{20}] |_{\theta=d=\theta_0} \\ + o(t^{-1}) \quad (3.2)$$

as $t \rightarrow \infty$, where

$$q_1 \triangleq \frac{\partial \log f_\theta}{\partial \theta}, \\ p_{12} \triangleq E_\theta \left\{ \frac{\partial \log f_{Y^t|\theta}}{\partial \theta} \frac{\partial^2 \log f_{Y^t|\theta}}{\partial \theta^2} \right\}, \\ p_1 \triangleq E_\theta \frac{\partial^2 \log f_{Y^t|\theta}}{\partial \theta^2}, \\ V_{i,s} \triangleq \frac{\partial^i}{\partial d^i} \frac{\partial^s}{\partial \theta^s} V(d, \theta).$$

See [5, p. 1147] for the derivation.

For the quadratic loss function, $V(d, \theta) = (d - \theta)^2$, it is easily verified that $\hat{\theta}_t$ is the posterior mean and $\arg \min_{\theta \in \Theta} E_{\theta_0} V(\hat{\theta}_t, \theta) = E_{\theta_0} \hat{\theta}_t$. Thus, in this special case, (3.2) gives an asymptotic expansion of the expected value of the posterior mean. The influence of the prior on the expected value is seen in the q_1 term.

From Lemma 3 we see that $p_2^{-1} = O(t^{-1})$ and the terms within the brackets of (3.2) are $O(1)$. Therefore, in general, $E_{\theta_0} \hat{\theta}_t = \theta_0 + O(t^{-1})$. Because $V_{12} = V_{21} = 0$, we can conclude that $E_{\theta_0} \hat{\theta}_t = \theta_0 + o(t^{-1})$ if and only if

$$[q_1 = (p_{12} + p_{31})/p_2]_{\theta=\theta_0}. \quad (3.3)$$

The prior that satisfies (3.3) is therefore the minimum bias prior.

Observe that

$$(p_{12} + p_{31})/p_2 = \frac{\partial}{\partial \theta} \log \left(-E_{\theta} \left\{ \frac{\partial^2 \log f_{Y|I|\theta}}{\partial \theta^2} \right\} \right)$$

upon interchanging the operations of differentiation and expectation (which is easily justified). Hence, a prior satisfying (3.3) is given by

$$f_{\theta} = -E_{\theta} \left\{ \frac{\partial^2 \log f_{Y|I|\theta}}{\partial \theta^2} \right\}. \quad (3.4)$$

Equation (3.4) is, for a scalar θ , a closed-form solution for the minimum bias prior. It is, in fact, the square of the Jeffreys prior. For the multivariate case, the situation is more complicated. Using the results [5, p. 1149] for a vector θ , where the loss function is given by $V(d, \theta) = \sum_{i=1}^p (d_i - \theta_i)^2$, $E_{\theta_0} \hat{\theta}_t = \theta_0 + o(t^{-1})$ if and only if

$$q_i = \sum_{j,k=1}^p (J^{-1})_{j,k} (p_{i,j,k} + p_{i,k,j}) \quad i = 1, \dots, p, \quad \theta = \theta_0 \quad (3.5)$$

where

$$\begin{aligned} p_{i,j,k} &\triangleq \frac{\partial}{\partial \theta_j} \log f_{\theta} \cdot \frac{\partial}{\partial \theta_k} \log f_{Y|I|\theta} \\ &\triangleq E_{\theta} \left\{ \frac{\partial^2 \log f_{Y|I|\theta}}{\partial \theta_i \partial \theta_j} \cdot \frac{\partial \log f_{Y|I|\theta}}{\partial \theta_k} \right\}, \\ p_{i,j,k} &\triangleq E_{\theta} \left\{ \frac{\partial^3 \log f_{Y|I|\theta}}{\partial \theta_i \partial \theta_j \partial \theta_k} \right\} \end{aligned}$$

and $J \triangleq E_{\theta} \{ \partial^2 \log f_{Y|I|\theta} / \partial \theta \partial \theta^T \}$. Equation (3.5) consists of p differential equations that the multivariate minimum bias prior must satisfy. We apply (3.5) to the model (2.1) and simplify the results in the following theorem.

Theorem 1: $E_{\theta_0} \hat{\theta}_t = \theta_0 + o(t^{-1})$ if and only if

$$h_i = \frac{1}{2} \frac{\partial}{\partial \theta_i} [\log \det(-J)] + \text{tr} \{ J^{-1} U_i \} \Big|_{\theta=\theta_0}, \quad i = 1, \dots, p \quad (3.6)$$

where

$$J = \frac{\partial l(\theta)}{\partial \theta} 2\Lambda_t \frac{\partial l(\theta)}{\partial \theta^T}, \quad (3.7)$$

$$(U_i)_{j,k} \triangleq \frac{\partial l^T(\theta)}{\partial \theta_i} 2\Lambda_t \frac{\partial^2 l(\theta)}{\partial \theta_j \partial \theta_k}, \quad j, k = 1, \dots, p \quad (3.8)$$

and Λ_t is defined recursively in (A.8b).

Proof: See Appendix D.

As it must, Theorem 1 yields (3.4) in the special case $p = 1$. To verify this, observe that

$$J = \frac{\partial l^T(\theta)}{\partial \theta} 2\Lambda_t \frac{\partial l(\theta)}{\partial \theta}, \quad U_1 = \frac{\partial l^T(\theta)}{\partial \theta} 2\Lambda_t \frac{\partial^2 l(\theta)}{\partial \theta^2}$$

and therefore $U_1 = \frac{1}{2} \partial J / \partial \theta$. It follows that $J^{-1} U_1 = \frac{1}{2} J^{-1} \partial J / \partial \theta = \frac{1}{2} \partial \log(-J) / \partial \theta$. Hence, from Theorem 1, the minimum bias prior satisfies $q_1 = \partial \log(-J) / \partial \theta$. The solution to this is (3.4).

The set of differential equations given by (3.6), with a closed-form solution for scalar θ , may not have a closed-form solution in the multivariate case for all θ . The problem, however, can be made tractable by observing that Theorem 1 requires the differential equations to hold only at θ_0 . θ_0 is unknown, but with an estimate we may construct a simple prior that satisfies the conditions of Theorem 1 approximately. This prior can then be used to obtain the next estimate of θ_0 , which in turn feeds the next prior, and so on, reducing the bias through an iterative procedure. This bootstrapping method is now stated and proved to have a beneficial effect on the bias.

Theorem 2. Define $\nu(\theta) = [\nu_1(\theta), \dots, \nu_p(\theta)]^T$, where

$$\nu_i(\theta) \triangleq \frac{1}{2} \frac{\partial}{\partial \theta_i} [\log \det(-J)] + \text{tr} \{ J^{-1} U_i \}, \quad i = 1, \dots, p. \quad (3.9)$$

Let Θ be a closed bounded set such that $\nu(\theta)$ is continuous for all $\theta \in \Theta$ and sufficiently large t . For every t , suppose some mechanism is used to "project" $\hat{\theta}_{t-1}$ onto Θ . Let $\hat{\theta}_{t-1}$ be this projection and

$$f_{\theta}(\hat{\theta}_{t-1}) = \exp \left(\theta^T \cdot \nu(\hat{\theta}_{t-1}) \right) \quad (3.10)$$

be the prior used at time t . Then, $E_{\theta_0} \hat{\theta}_t = \theta_0 + o(t^{-1})$ as $t \rightarrow \infty$.

Proof: See Appendix E.

This provides a practical method for achieving the minimum bias effect without having to explicitly solve for the minimum bias prior. The prior defined in (3.10) has a simple form and is readily calculated at any time t . It is adaptive since it depends on $\hat{\theta}_{t-1}$ and therefore also Y^{t-1} .

Remarks: The projection mechanism should modify only estimates that fall outside Θ by, say, assigning them boundary values of Θ . The actual mechanism will not be important because θ_0 is an interior point of Θ , and given enough observations, $|\hat{\theta}_t - \theta_0|$ will be large with only small probability. Hence the projection mechanism will generally modify the estimates only when t is small. The mechanism's main use is

to provide a deterministic bound for the random variable $\tilde{\theta}_t$, and to ensure that the prior given by (3.10) is always well defined.

The choice of initial prior (before any estimate of θ_0 is available) is arbitrary and will not affect the convergence behavior of $E_{\theta_0} \tilde{\theta}_t$ for large enough t . A convenient choice that generally provides good initial parameter estimates is the uniform density, that is $f_{\theta} = c$ for some arbitrary constant $c > 0$. In fact, a good rule of thumb is to employ the uniform density until the terms neglected in the asymptotic expansions can safely be ignored. At this point, the adaptive prior may be employed. Otherwise, if (3.10) is used too early, erratic estimate behavior could result. That is, a poor estimate of θ_0 used to construct $f_{\theta}(\tilde{\theta}_{t-1})$ may lead to a possibly worse estimate, and so on. Nevertheless, even this process will eventually correct itself as t becomes larger. The smallest t for which the minimum bias prior should be applied can be determined by experimentation.

IV. EXAMPLES AND SIMULATIONS

We give two examples, one that has a closed-form solution for the minimum bias prior and one that does not.

Example 1—Linear in the Parameters: Let $g(\cdot)$ and $h(\cdot)$ be linear so $l(\theta) = \theta$. The posterior density of the parameters is given by Lemma 1

$$f_{\theta|Y} \propto f_{\theta} \exp \{ \alpha_t^T \theta + \theta^T \Lambda_t \theta \}.$$

Calculating the minimum bias prior using Theorem 1 yields $J = 2\Lambda_t$, $U_t = 0$, $t = 1, \dots, p$. Therefore, the prior satisfies $q_t = 0$, $t = 1, \dots, p$. A closed-form solution is given by the uniform density $f_{\theta} = c$ for some arbitrary $c > 0$. The posterior mean, when $f_{\theta} = c$, is $\tilde{\theta}_t = -(1/2)\Lambda_t^{-1}\alpha_t$, since $f_{\theta|Y} \propto \exp \{ \alpha_t^T \theta + \theta^T \Lambda_t \theta \}$ is a Gaussian distribution. It follows from (A.8) that $E_{\theta_0} \tilde{\theta}_t = \theta_0$, and it is straightforward to verify that using the prior $f_{\theta} = c$ is equivalent to employing (2.2) and initializing the inverse covariance of θ to zero.

This example shows that when g and h are linear, closed-form solutions for $\tilde{\theta}_t$ and the minimum bias prior are available at every time step. The minimum bias prior, in this case, yields a completely unbiased estimate for all t . We next give an example with nonlinear g in which the adaptive prior must be employed to minimize bias.

Example 2—Determining the Thrust of an Accelerating Target: We look at the problem of estimating the thrust of a moving target in two dimensions, where we are allowed vector measurements of the x and y -axis coordinate positions and velocities. The target is assumed to have a fixed unknown thrust a along an unknown direction φ during the observation interval. The direction is taken with respect to the y -axis and the observation intervals have length T . The velocity obeys

$$\begin{aligned} \begin{bmatrix} u_x(t) \\ u_y(t) \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_x(t-1) \\ u_y(t-1) \end{bmatrix} \\ &+ \begin{bmatrix} Ta \sin \varphi \\ T[a \cos \varphi - g] \end{bmatrix} + w_t' \end{aligned} \quad (4.1)$$

where

$$E w_t' w_t'^T = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

and g is gravitational acceleration. The unequal noise variances are due to the greater atmospheric disturbance parallel to the x -axis. The equations of position $r(t)$ are assumed to obey

$$r(t) = r(t-1) + \frac{t}{2} [u(t-1) + u(t)]. \quad (4.2)$$

Letting $x_t \triangleq [u_x(t), u_y(t), r_x(t), r_y(t)]^T$ and combining (4.1) and (4.2) we see that

$$\begin{aligned} x_t &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ T & 0 & 1 & 0 \\ 0 & T & 0 & 1 \end{bmatrix} x_{t-1} \\ &+ \begin{bmatrix} Ta \sin \varphi \\ T[a \cos \varphi - g] \\ (T^2/2)a \sin \varphi \\ (T^2/2)[a \cos \varphi - g] \end{bmatrix} + w_{t-1} \end{aligned} \quad (4.3)$$

and

$$E w_t w_t^T = \begin{bmatrix} 0 & T & 0 \\ T & 0 & T/2 \\ 0 & T/2 & 0 \\ 0 & T/2 & 0 & T^2/4 \end{bmatrix}$$

We assume the observations follow the model $y_t = C'x_t + v_t$, where $E v_t v_t^T = I$ and C' describes the sensors' outputs as a function of target velocity and position.

We wish to estimate $\theta = [a, \varphi]^T$ using the minimum bias prior. For simplicity we suppose that $C'^T C' = I$, $m_0 = 0$ and $P_0 = I$ [see Assumption A1]. We have

$$\begin{aligned} B &= \begin{bmatrix} T & 0 \\ 0 & T \\ T^2/2 & 0 \\ 0 & T^2/2 \end{bmatrix} \\ g(\theta) &= [a \sin \varphi, a \cos \varphi - g]^T, \\ h(\theta) &\equiv 0 \end{aligned}$$

From (3.6)

$$\frac{\partial \log f_{\theta}}{\partial a} = \frac{\xi_t - 1}{a \xi_t} [1 - 2\xi_t \cos^2 \varphi + (\xi_t - 1) \cos^4 \varphi], \quad (4.4a)$$

$$\frac{\partial \log f_{\theta}}{\partial \varphi} = \frac{\xi_t - 1}{\xi_t} [1 + 2\xi_t + (\xi_t - 1) \cos^2 \varphi] \sin \varphi \cos \varphi \quad (4.4b)$$

where $\xi_t \triangleq (\Lambda_t)_{2,2}/(\Lambda_t)_{1,1}$. We do not know of a closed-form expression for f_{θ} satisfying these equations.

Instead, we use the prior density of Theorem 2. Then, at time t

$$f_{\theta}(\tilde{\theta}_{t-1}) = \exp \left(a \nu_1(\tilde{\theta}_{t-1}) + \varphi \nu_2(\tilde{\theta}_{t-1}) \right)$$

where $\nu_1(\tilde{\theta}_{t-1})$ and $\nu_2(\tilde{\theta}_{t-1})$ are the right-hand sides of, respectively, (4.4a) and (4.4b) evaluated at $a = \tilde{a}_{t-1}$ and

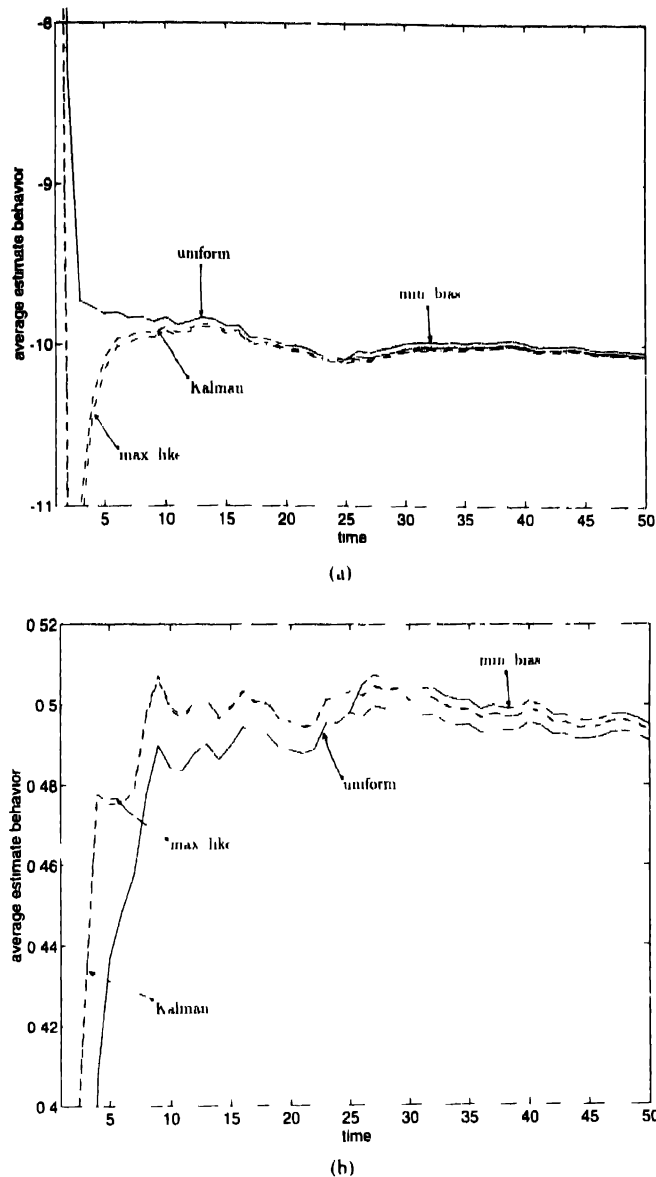


Fig. 1 Posterior mean estimates computed with adaptive minimum bias and uniform priors, averaged over 100 trials. The minimum bias prior was applied when $t \geq 25$. Also shown are the extended-state Kalman filter and maximum likelihood estimates. Solid lines correspond to minimum bias and uniform priors, dashed to Kalman filter, dash-dotted to maximum likelihood (a) Estimates of $a_0 = -10$ (b) Estimates of $\varphi_0 = 0.5$

$\varphi = \tilde{\varphi}_{t-1}$. Observe that from (4.4), $\nu(\theta)$ is continuous except when $a = 0$. To fulfill the conditions of Theorem 2 it therefore suffices to let $\Theta = \{\theta \in \mathbb{R}^2 \mid \delta \leq |a| \leq 50, \varphi \in [-\pi, \pi)\}$ where δ is any small positive number. This imposes a weak and practical restriction on the choice of parameters since the choice $a = 0$ is not identifiable in φ and therefore does not satisfy Assumption B3) anyway.

We simulated the system with $T = 0.25$, $a_0 = -10$, and $\varphi_0 = 0.5$. The posterior mean estimates with the minimum bias and the uniform priors and, for the sake of comparison, the extended-state Kalman filter and the maximum likelihood estimates are all shown in Fig. 1. Plotted are the estimates' responses averaged over 100 trials; the averages approximate the expected curve values. We show the first 50 out of the total of 250 time steps to demonstrate the effect of application

TABLE I
BIAS PERFORMANCE OF VARIOUS ESTIMATES

Estimate	Deviation around	
	a_0	φ_0
min. bias	5.23	1.03
uniform	6.80	1.41
Kalman	7.45	1.16
max. like.	7.92	1.15

of the minimum bias prior at $t = 25$. The prior was uniform for $t < 25$ (see remarks after Theorem 2). The extended-state Kalman filter used the fact that the state-space model is linear in $g(\theta)$ to get estimates of $a \sin \varphi$ and $a \cos \varphi - g$, which were converted into estimates of a and φ . The $g(\theta)$ portion of the initial covariance matrix was set to $10^3 I$ to approximately convey ignorance, and the initial mean was $[0, -\bar{g}]^T$. The maximum likelihood estimate was evaluated in the usual way.

Table I is a measure of the bias of each method. We show the magnitude of the difference from the true parameter of each curve summed from time steps 25–250. From the table it is clear that the posterior mean using the minimum bias prior has the least deviation of all the estimates. That the minimum bias prior causes less deviation than the uniform prior is predicted by the theory. That it happens to cause less deviation than the Kalman filter and maximum likelihood estimates may be just fortuitous. Nevertheless, this example demonstrates the viability of the minimum bias prior.

V. CONCLUSION

The minimum bias prior was proposed as an alternative to other prior densities based on ignorance because of its beneficial effect on the posterior mean. Within a realization, for t sufficiently large, the dominant effect of any prior appeared as bias. The minimum bias prior ensured that $E_{\theta_0} \hat{\theta}_t - \theta_0 = o(t^{-1})$, as $t \rightarrow \infty$, as opposed to $O(t^{-1})$ for other priors.

In Example 2, the proposed estimate was shown to have bias performance superior to the maximum likelihood and Kalman filter estimates. Some additional tests were performed indicating that this improvement held for other parameter values as well. It would be of interest to determine analytically if this is a general phenomenon.

When the bias contribution to the realization mean-square error, $E_{\theta_0}(\hat{\theta}_t - \theta_0)^2$, is significant, we can expect the minimum bias prior to reduce this error. We performed some numerical experiments to see if $E_{\theta_0}(\hat{\theta}_t - \theta_0)^2$ was lowest for the minimum bias estimate in Example 2, but the realization mean-square errors of all the estimates in Table I were essentially equal. We discovered that this was due to the overwhelming contribution of the variance over the bias to the realization mean-square error in this example. This, of course, raises the question: Is there a systematic way to choose priors to minimize this error?

Bayes estimates other than the posterior mean can be analyzed by choosing nonquadratic V in the asymptotic expansion (3.2). The minimum bias prior and its adaptive version vary

with V , and so do their interpretations. This is a potential subject for further investigation.

APPENDIX A PROOF OF LEMMA 1

The notion of recursively calculating density functions in state-space models can be found, for example, in [3] and the proof of Lemma 1 follows the same, now standard, ideas. Due to the structure of the state-space model (2.1), the recursive scheme for computing the posterior density function is particularly simple.

Given θ , (2.1) obeys a standard linear state-space model with known additive terms, implying that

$$f_{x_t|x_{t-1}, \theta} = \mathcal{N}(Ax_{t-1} + Bg(\theta), Q),$$

$$f_{y_t|x_t, \theta} = \mathcal{N}(Cx_t + Dh(\theta), R).$$

Using Assumptions A1)–A2), given θ and Y^t we observe that the density of x_t is Gaussian. That is

$$f_{x_t|Y^t, \theta} = \mathcal{N}(m_t, P_t) \quad (\text{A.1})$$

where m_t and P_t may be computed iteratively. The equations for the recursion are given by the Kalman filter:

- i) $m_t = Am_{t-1} + Bg(\theta) + K_t[y_t - C(Am_{t-1} + Bg(\theta)) - Dh(\theta)]$,
- ii) $K_t = (AP_{t-1}A^T + Q)C^T S_t^{-1}$,
- iii) $P_t = (I - K_t C)(AP_{t-1}A^T + Q)$,

where

$$S_t \triangleq C(AP_{t-1}A^T + Q)C^T + R \quad (\text{A.2})$$

for $t = 1, 2, \dots$. By inspecting i), we conclude that

$$m_t = M_t l(\theta) + z_t \quad (\text{A.3})$$

where

$$z_t = (I - K_t C)Az_{t-1} + K_t y_t, \quad (\text{A.4a})$$

$$M_t = (I - K_t C)AM_{t-1} + [(I - K_t C)B, -K_t D] \quad (\text{A.4b})$$

with M_0 equal to zero and $z_0 = m_0$. This gives (2.4).

Plainly, $f_{x_t, \theta|Y^t} = f_{x_t|Y^t, \theta} \cdot f_{\theta|Y^t}$, which is (2.3). We have dispensed with $f_{x_t|Y^t, \theta}$, so we need only to evaluate $f_{\theta|Y^t}$. By a standard argument

$$f_{\theta|Y^t} \propto f_{\theta} \cdot f_{Y^t|\theta} = f_{\theta} \cdot \prod_{r=1}^t f_{y_r|Y^{r-1}, \theta} \quad (\text{A.5})$$

and it follows from (2.1) and (A.1) that

$$f_{y_t|Y^{t-1}, \theta} = \mathcal{N}(N_t l(\theta) + CAz_{t-1}, S_t) \quad (\text{A.6})$$

where

$$N_t \triangleq CAM_{t-1} + [CB, D]. \quad (\text{A.7})$$

We may now prove (2.5) by induction. At $t = 0$, since Y^0 is the empty set, we set α_0 and Λ_0 to zero and (2.5) holds.

Now assume (2.5) holds at time $t-1 \geq 0$. We need to show that it also holds at time t . Equations (A.5) and (A.6) yield

$$f_{\theta|Y^t} \propto f_{\theta|Y^{t-1}} \cdot f_{y_t|Y^{t-1}, \theta}$$

$$\propto f_{\theta} \cdot \exp \{ \alpha_{t-1}^T l(\theta) + l(\theta)^T \Lambda_{t-1} l(\theta) \}$$

$$\cdot \mathcal{N}(y_t; N_t l(\theta) + CAz_{t-1}, S_t).$$

It is straightforward to find, upon substituting (A.3) for m_t , that the posterior density retains its form. The iterative equations are

$$\alpha_t = \alpha_{t-1} + N_t^T S_t^{-1} (y_t - CAz_{t-1}), \quad (\text{A.8a})$$

$$\Lambda_t = \Lambda_{t-1} - \frac{1}{2} N_t^T S_t^{-1} N_t \quad (\text{A.8b})$$

for $t = 1, 2, \dots$.

APPENDIX B PROOF OF LEMMA 3

The proof of part i) uses a conditioning argument to show that certain random variables, of which η_s and η_t are functions, are independent when $s \neq t$. This will imply the independence of η_s and η_t . The proof of part ii) is a simple application of Lemma 2.

i) Define $\varepsilon_t(\theta, Y^t) = y_t - N_t l(\theta) - CAz_{t-1}$, and assume that θ is given. Clearly $\varepsilon_t(\theta, Y^t)$ is a linear combination of Gaussian random variables and therefore also has a Gaussian distribution. Letting $\Delta_t(\theta_0, \theta) \triangleq N_t[l(\theta_0) - l(\theta)]$, we obtain

$$E_{\theta_0} \varepsilon_t(\theta, Y^t) = E_{\theta_0} \{ E_{\theta_0} \varepsilon_t(\theta, Y^t) | Y^{t-1} \}$$

$$= E_{\theta_0} \{ N_t [l(\theta_0) - l(\theta)] \}$$

$$= \Delta_t(\theta_0, \theta) \quad (\text{B.1})$$

because $f_{y_t|Y^{t-1}, \theta_0} = \mathcal{N}(N_t l(\theta_0) + CAz_{t-1}, S_t)$. Furthermore, for $s < t$

$$E_{\theta_0} \varepsilon_s(\theta, Y^s) \varepsilon_t^T(\theta, Y^t) = E_{\theta_0} \{ E_{\theta_0} \varepsilon_s(\theta, Y^s) \varepsilon_t^T(\theta, Y^t) | Y^{t-1} \}$$

$$= E_{\theta_0} \{ \varepsilon_s(\theta, Y^s) E_{\theta_0} \varepsilon_t^T(\theta, Y^t) | Y^{t-1} \}$$

$$= [E_{\theta_0} \varepsilon_s(\theta, Y^s)] \Delta_t^T(\theta_0, \theta)$$

$$= \Delta_s(\theta_0, \theta) \Delta_t^T(\theta_0, \theta) \quad (\text{B.2})$$

so ε_s and ε_t are uncorrelated. We show that they are jointly Gaussian and therefore also independent; it follows that $\eta_s = -\frac{1}{2} [\log \det 2\pi S_s + \varepsilon_s^T S_s^{-1} \varepsilon_s]$ and $\eta_t = -\frac{1}{2} [\log \det 2\pi S_t + \varepsilon_t^T S_t^{-1} \varepsilon_t]$ are independent as well.

We have $f_{\varepsilon_t, \varepsilon_s | \theta_0} = f_{\varepsilon_t | \varepsilon_s, \theta_0} \cdot f_{\varepsilon_s | \theta_0}$, where $f_{\varepsilon_s | \theta_0}$ is Gaussian. Moreover, $f_{\varepsilon_t | \varepsilon_s, \theta_0} = \int f_{\varepsilon_t | Y^{t-1}, \varepsilon_s, \theta_0} \cdot f_{Y^{t-1} | \varepsilon_s, \theta_0} dY^{t-1} = \int f_{\varepsilon_t | Y^{t-1}, \theta_0} \cdot f_{Y^{t-1} | \varepsilon_s, \theta_0} dY^{t-1}$. From (B.1) and the distribution of y_t , $f_{\varepsilon_t | Y^{t-1}, \theta_0} = \mathcal{N}(\Delta_t(\theta_0, \theta), S_t)$, and because ε_s is a linear combination of Y^s , $f_{Y^{t-1} | \varepsilon_s, \theta_0}$ is also Gaussian. Therefore, $f_{\varepsilon_t | \varepsilon_s, \theta_0}$ is Gaussian, and so is $f_{\varepsilon_t, \varepsilon_s | \theta_0}$.

ii) The random variable $S_t^{-1/2} \varepsilon_t(\theta, Y^t)$ is Gaussian with mean $S_t^{-1/2} \Delta_t(\theta_0, \theta)$ and covariance I . According to Lemma 2, $\Delta_t(\theta_0, \theta) \rightarrow \Delta_\infty(\theta_0, \theta) = N_\infty[l(\theta_0) - l(\theta)]$ and $S_t \rightarrow S_\infty$, as $t \rightarrow \infty$. Consequently, $S_t^{-1/2} \varepsilon_t(\theta, Y^t) \rightarrow S_\infty^{-1/2} \varepsilon_\infty$ in distribution. Since η_t depends continuously on $S_t^{-1/2} \varepsilon_t(\theta, Y^t)$, we have $\eta_t \rightarrow \eta_\infty$ in distribution as well.

To show convergence of the moments of the derivatives of η_t we give two examples: η_t itself and the square of its first

derivative. All of the remaining cases are shown in exactly the same way. For η_t we have

$$\begin{aligned} E_{\theta_0} \eta_t(\theta) &= -\frac{1}{2} [\log \det 2\pi S_t + \Delta_t^T S_t^{-1} \Delta_t \\ &\quad + \text{tr } S_t^{-1} E_{\theta_0}(\varepsilon_t - \Delta_t)(\varepsilon_t - \Delta_t)^T \\ &\quad + 2\Delta_t^T S_t^{-1} E_{\theta_0}(\varepsilon_t - \Delta_t)] \\ &= -\frac{1}{2} [\log \det 2\pi S_t + \Delta_t^T S_t^{-1} \Delta_t + m]. \quad (\text{B.3}) \end{aligned}$$

Similarly, $E \eta_\infty = -\frac{1}{2} [\log \det 2\pi S_\infty + \Delta_\infty^T S_\infty^{-1} \Delta_\infty + m]$. The result now follows from Lemma 2. For the square of the first derivative

$$\begin{aligned} E_{\theta_0} \frac{\partial \eta_t(\theta)}{\partial \theta_i} \frac{\partial \eta_t(\theta)}{\partial \theta_j} &= E_{\theta_0} \frac{\partial \varepsilon_t^T}{\partial \theta_i} S_t^{-1} \varepsilon_t \varepsilon_t^T S_t^{-1} \frac{\partial \varepsilon_t}{\partial \theta_j} \\ &\quad + \frac{\partial l^T(\theta)}{\partial \theta_i} N_t^T S_t^{-1} E_{\theta_0} \varepsilon_t \varepsilon_t^T \\ &\quad + S_t^{-1} N_t \frac{\partial l(\theta)}{\partial \theta_j} \\ &= \frac{\partial l^T(\theta)}{\partial \theta_i} N_t^T S_t^{-1} (S_t + \Delta_t \Delta_t^T) \\ &\quad + S_t^{-1} N_t \frac{\partial l(\theta)}{\partial \theta_j}. \end{aligned}$$

Similarly

$$\begin{aligned} E \frac{\partial \eta_\infty(\theta)}{\partial \theta_i} \frac{\partial \eta_\infty(\theta)}{\partial \theta_j} &= \frac{\partial l^T(\theta)}{\partial \theta_i} N_\infty^T S_\infty^{-1} \\ &\quad + (S_\infty + \Delta_\infty \Delta_\infty^T) S_\infty^{-1} N_\infty \frac{\partial l(\theta)}{\partial \theta_j} \end{aligned}$$

and the result again follows from Lemma 2.

APPENDIX C PROOF OF LEMMA 4

This lemma is proved here using Chebyshev's inequality in much the same way as Lemma 1 of [5]. Our proof, however, accommodates the fact that $\log f_{1:\theta}$ is a sum of independent random variables that do not have identical distributions.

We start with $\log(f_{1:\theta}/f_{1:\theta_0}) = \sum_{r=1}^t \{\eta_r(\theta) - \eta_r(\theta_0)\}$, where $\eta_r(\theta)$ is defined in (3.1). From (B.3)

$$\begin{aligned} E_{\theta_0} \{\eta_t(\theta) - \eta_t(\theta_0)\} &= -\frac{1}{2} \Delta_t^T(\theta_0, \theta) S_t^{-1} \Delta_t(\theta_0, \theta) \\ &\quad + \frac{1}{2} \Delta_t^T(\theta_0, \theta_0) S_t^{-1} \Delta_t(\theta_0, \theta_0) \\ &= -\frac{1}{2} \Delta_t^T(\theta_0, \theta) S_t^{-1} \Delta_t(\theta_0, \theta). \end{aligned}$$

Therefore, Assumption B3) and Lemma 3 imply that $E_{\theta_0} \{\eta_t(\theta) - \eta_t(\theta_0)\}$ tends exponentially quickly, as $t \rightarrow \infty$, to some $\mu_1(\theta) < 0$ when $\theta \neq \theta_0$. In general, $\eta_r(\theta) - \eta_r(\theta_0)$ has s th central moment converging exponentially quickly to some $\mu_s(\theta)$. The moments are continuous functions of θ , so if we let $\mu_s \triangleq \max_{\epsilon < |\theta - \theta_0| \leq k} \mu_s(\theta)$ for any $0 < \epsilon < k$, then

for any θ obeying $\epsilon \leq |\theta - \theta_0| \leq k$

$$\begin{aligned} &P \left\{ \sum_{r=1}^t \eta_r(\theta) - \eta_r(\theta_0) > \frac{1}{2} t \mu_1 + O(1) \right\} \\ &\leq P \left\{ \sum_{r=1}^t \eta_r(\theta) - \eta_r(\theta_0) > \frac{1}{2} t \mu_1(\theta) + O(1) \right\} \\ &\leq P \left\{ \left| \sum_{r=1}^t [\eta_r(\theta) - \eta_r(\theta_0)] - t \mu_1(\theta) \right| > -\frac{1}{2} t \mu_1(\theta) \right\} \\ &\leq \frac{3t(t-1)\mu_2^2(\theta) + t\mu_4(\theta) + O(1)}{(1/16)t^4\mu_1^4(\theta)} \\ &\leq \frac{3t(t-1)\mu_2^2 + t\mu_4 + O(1)}{(1/16)t^4\mu_1^4} = O(t^{-2}) \end{aligned}$$

as $t \rightarrow \infty$. Equivalently, $f_{Y^{1:\theta}}/f_{Y^{1:\theta_0}}$ is $O(e^{-t\delta})$ with probability $1 - O(t^{-2})$. The constants in the $O(\cdot)$ terms are independent of θ so the approximation is uniform. From Assumption B6), the result holds for $|\theta| > k$ when k is sufficiently large. We therefore have uniformity in $|\theta - \theta_0| > \epsilon$ for any $\epsilon > 0$.

APPENDIX D PROOF OF THEOREM 1

We prove Theorem 1 by evaluating the expressions of (3.5) and simplifying. From (A.6) we obtain

$$f_{1:\theta} = \prod_{r=1}^t \mathcal{N}(y_r; N_r l(\theta) + C A z_{r-1}, S_r).$$

So

$$\frac{\partial \log f_{1:\theta}}{\partial \theta_i} = \sum_{r=1}^t \frac{\partial l^T(\theta)}{\partial \theta_i} N_r^T S_r^{-1} \varepsilon_r,$$

where $\varepsilon_r(\theta, Y^r) = y_r - N_r l(\theta) - C A z_{r-1}$. Also

$$\begin{aligned} \frac{\partial^2 \log f_{1:\theta}}{\partial \theta_i \partial \theta_j} &= \sum_{r=1}^t \left[\frac{\partial^2 l^T(\theta)}{\partial \theta_i \partial \theta_j} N_r^T S_r^{-1} \varepsilon_r \right. \\ &\quad \left. - \frac{\partial l^T(\theta)}{\partial \theta_i} N_r^T S_r^{-1} N_r \frac{\partial l(\theta)}{\partial \theta_j} \right] \\ \frac{\partial^3 \log f_{1:\theta}}{\partial \theta_i \partial \theta_j \partial \theta_k} &= \sum_{r=1}^t \left[\frac{\partial^3 l^T(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k} N_r^T S_r^{-1} \varepsilon_r \right. \\ &\quad - \frac{\partial^2 l^T(\theta)}{\partial \theta_i \partial \theta_j} N_r^T S_r^{-1} N_r \frac{\partial l(\theta)}{\partial \theta_k} \\ &\quad - \frac{\partial l^T(\theta)}{\partial \theta_i} N_r^T S_r^{-1} N_r \frac{\partial^2 l(\theta)}{\partial \theta_j \partial \theta_k} \\ &\quad \left. - \frac{\partial^2 l^T(\theta)}{\partial \theta_i \partial \theta_k} N_r^T S_r^{-1} N_r \frac{\partial l(\theta)}{\partial \theta_j} \right]. \end{aligned}$$

Therefore

$$\begin{aligned} (J)_{j,k} &= - \sum_{r=1}^t \frac{\partial l^T(\theta)}{\partial \theta_j} N_r^T S_r^{-1} N_r \frac{\partial l(\theta)}{\partial \theta_k} \\ &= \frac{\partial l^T(\theta)}{\partial \theta_j} 2\Lambda_t \frac{\partial l(\theta)}{\partial \theta_k} \quad (\text{D.1}) \end{aligned}$$

where the last equality is a consequence of (A.8b). Furthermore

$$p_{ijk} = E_{\theta} \left\{ \sum_{r=1}^t \left[\frac{\partial^2 l^T(\theta)}{\partial \theta_i \partial \theta_j} N_r^T S_r^{-1} \right. \right. \\ \left. \left. - \frac{\partial l^T(\theta)}{\partial \theta_i} N_r^T S_r^{-1} N_r \frac{\partial l(\theta)}{\partial \theta_j} \right] \right. \\ \left. \cdot \sum_{s=1}^t \frac{\partial l^T(\theta)}{\partial \theta_k} N_s^T S_s^{-1} \varepsilon_s \right\}. \quad (D.2)$$

From (B.1) and (B.2), $E_{\theta} \varepsilon_t = \Delta_t(\theta, \theta) = 0$; whenever $s \neq t$, $E_{\theta} \varepsilon_s \varepsilon_t^T = \Delta_s(\theta, \theta) \Delta_t^T(\theta, \theta) = 0$; $E_{\theta} \varepsilon_t \varepsilon_t^T = S_t + \Delta_t(\theta, \theta) \Delta_t^T(\theta, \theta) = S_t$. It follows that

$$p_{ijk} = - \frac{\partial^2 l^T(\theta)}{\partial \theta_i \partial \theta_j} 2\Lambda_t \frac{\partial l(\theta)}{\partial \theta_k} \quad (D.3)$$

Similarly

$$p_{ijk} = \frac{\partial^2 l^T(\theta)}{\partial \theta_i \partial \theta_j} 2\Lambda_t \frac{\partial l(\theta)}{\partial \theta_k} + \frac{\partial^2 l^T(\theta)}{\partial \theta_j \partial \theta_k} 2\Lambda_t \frac{\partial l(\theta)}{\partial \theta_i} \\ + \frac{\partial^2 l^T(\theta)}{\partial \theta_i \partial \theta_k} 2\Lambda_t \frac{\partial l(\theta)}{\partial \theta_j}. \quad (D.4)$$

Combining (D.3) and (D.4), we obtain

$$p_{ijk} + p_{ikj} = \frac{\partial^2 l^T(\theta)}{\partial \theta_i \partial \theta_k} 2\Lambda_t \frac{\partial l(\theta)}{\partial \theta_j} + \frac{\partial^2 l^T(\theta)}{\partial \theta_j \partial \theta_k} 2\Lambda_t \frac{\partial l(\theta)}{\partial \theta_i}$$

and using (3.5) and (D.1), we see that the minimum bias prior must satisfy

$$= \sum_{j,k=1}^p (J^{-1})_{j,k} \\ - \frac{1}{2} \frac{\partial}{\partial \theta_i} \{ (J)_{j,k} \} + \frac{\partial l^T(\theta)}{\partial \theta_i} 2\Lambda_t \frac{\partial^2 l(\theta)}{\partial \theta_j \partial \theta_k} \\ - \frac{1}{2} \left\{ \text{tr} \left\{ J^{-1} \frac{\partial J}{\partial \theta_i} \right\} + \text{tr} \{ J^{-1} U_i \} \right\}_{\theta=\theta_0} \\ = \left[\frac{1}{2} \frac{\partial}{\partial \theta_i} [\log \det(-J)] + \text{tr} \{ J^{-1} U_i \} \right]_{\theta=\theta_0}, \\ i = 1, \dots, p.$$

This concludes the proof.

APPENDIX E PROOF OF THEOREM 2

We prove the result for a scalar θ ; the extension to the multivariate case is straightforward.

We first verify that the prior defined in (3.10) satisfies Assumptions B4) and B5). To verify B4) observe that $\log f_{\theta}(\hat{\theta}_{t-1}) = \theta \cdot \nu(\hat{\theta}_{t-1})$, and the derivatives $(\partial/\partial\theta) \log f_{\theta}(\hat{\theta}_{t-1}) = \nu(\hat{\theta}_{t-1})$, $(\partial^2/\partial\theta^2) \log f_{\theta}(\hat{\theta}_{t-1}) = 0$, exist and are continuous for all θ . For $\theta \in \Theta$, it follows from (A.6) and Lemma 2 that $f_{\theta}(\hat{\theta}_{t-1})$ is bounded as a function of y_t , uniformly in Y^{t-1} and t . Because Λ_t/t converges as $t \rightarrow \infty$ and Θ is bounded, $\nu(\theta)$ is continuous in θ uniformly in (sufficiently large) t . Thus, $f_{\theta}(\hat{\theta}_{t-1})$ is bounded and B5) is satisfied.

The fact that the prior is now a function of the observations modifies the asymptotic expansion (3.2). Since $\hat{\theta}_{t-1} \in \Theta$ and

$\nu(\theta)$ is continuous on Θ uniformly in t , we are assured that $q_1 = (\partial/\partial\theta) \log f_{\theta}(\hat{\theta}_{t-1}) = \nu(\hat{\theta}_{t-1})$ is still $O(1)$. Hence, the expansion (3.2) becomes

$$E_{\theta_0} \hat{\theta}_t = \theta_0 - p_2^{-1} [E_{\theta_0} q_1 - (p_{12} + p_{33})/p_2]_{\theta=\theta_0} \\ + o(t^{-1}) \quad (E.1)$$

as $t \rightarrow \infty$. It remains to show that $[E_{\theta_0} q_1 - (p_{12} + p_{33})/p_2]_{\theta=\theta_0} = o(1)$.

From [5, Theorem 2], $\hat{\theta}_t$ converges in probability to θ_0 , and since θ_0 is an interior point of Θ , $\hat{\theta}_t$ converges in probability to θ_0 (the projection mechanism has no effect in a small enough neighborhood of θ_0 ; see remarks after statement of theorem). For any $\delta > 0$ we have

$$|E_{\theta_0} q_1 - (p_{12} + p_{33})/p_2|_{\theta=\theta_0} = |E_{\theta_0} \nu(\hat{\theta}_{t-1}) - \nu(\theta_0)| \\ \leq E_{\theta_0} 1_{\{|\hat{\theta}_{t-1} - \theta_0| < \delta\}} |\nu(\hat{\theta}_{t-1}) - \nu(\theta_0)| \\ + E_{\theta_0} 1_{\{|\hat{\theta}_{t-1} - \theta_0| \geq \delta\}} |\nu(\hat{\theta}_{t-1}) - \nu(\theta_0)| \quad (E.2)$$

Let $\epsilon > 0$ be chosen. Because $\nu(\theta)$ is continuous at θ_0 , the first term on the right-hand side of (E.2) is bounded by ϵ for δ small enough. We know $\hat{\theta}_{t-1} \in \Theta$ is bounded and $\nu(\theta)$ is continuous for $\theta \in \Theta$, so the second term is bounded by a constant multiple of $P\{|\hat{\theta}_{t-1} - \theta_0| \geq \delta\}$. For t sufficiently large this term is also bounded by ϵ . Thus, $[E_{\theta_0} q_1 - (p_{12} + p_{33})/p_2]_{\theta=\theta_0} = o(1)$ as $t \rightarrow \infty$, and $E_{\theta_0} \hat{\theta}_t = \theta_0 - p_2^{-1} o(1) + o(t^{-1}) = \theta_0 + o(t^{-1})$, which was to be shown.

ACKNOWLEDGMENT

The authors thank Profs. J. Hartigan at Yale University and B. Porat at the Technion for helpful discussions. The authors are also grateful to the referees for their constructive comments.

REFERENCES

- [1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Englewood Cliffs, NJ: Prentice-Hall, 1979.
- [2] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag, 1985.
- [3] H. Cox, "On the estimation of state variables and parameters for noisy dynamic systems," *IEEE Trans Automat Contr*, vol. AC-9, pp. 5-12, Jan. 1964.
- [4] B. Friedland, "Treatment of bias in recursive filtering," *IEEE Trans Automat Contr*, vol. 14, pp. 359-367, Aug. 1969.
- [5] J. A. Hartigan, "The asymptotically unbiased prior distribution," *Ann Math Statistics*, vol. 36-4, pp. 1137-1152, Aug. 1965.
- [6] S. D. Hill and J. C. Spall, "Shannon information-theoretic priors for state-space model parameters," in *Bayesian Analysis of Time Series and Dynamic Models*, J. C. Spall, Ed., New York: Marcel Dekker, 1988, pp. 509-524.
- [7] H. Jeffreys, *Theory of Probability*, 3rd ed., Oxford: Clarendon Press, 1983.
- [8] J. Lin and A. P. Sage, "Algorithms for discrete sequential maximum likelihood bias estimation and associated error analysis," *IEEE Trans Sys Man Cybern*, vol. 1, pp. 314-324, Oct. 1971.
- [9] J. C. Spall and S. D. Hill, "Least-informative Bayesian prior distributions for finite samples based on information theory," *IEEE Trans Automat Contr*, vol. 35, pp. 580-583, May 1990.



Bertrand Hochwald (S'90) was born in New York, NY. He received the undergraduate education from Swarthmore College, Swarthmore, PA and the M.S. degree in electrical engineering from Duke University, Durham, NC. In 1989 he enrolled at Yale University, New Haven, CT, where he received the M.A. degree in statistics and is currently pursuing the Ph.D. degree in electrical engineering.

From 1986 to 1989 he worked as a design engineer for the Department of Defense at Fort Meade, MD. His research interests include statistical signal

processing, probability theory, and information theory.

Mr. Hochwald is the recipient of several achievement awards while employed at the Department of Defense and the Prize Teaching Fellowship at Yale. He is a member of Eta Kappa Nu.



Arye Nehorai (S'80, M'83, SM'90–F'94) received the B.Sc. and the M.Sc. degrees in electrical engineering from the Technion—Israel Institute of Technology in 1976 and 1979, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1983.

From 1983 to 1984 he was a Research Associate at Stanford University. From 1984 to 1985 he was a Research Engineer at Systems Control Technology, Inc., in Palo Alto, CA. Since 1985 he has been with the Department of Electrical Engineering at Yale

University, New Haven, CT, where he is an Associate Professor.

Dr. Nehorai is an Associate Editor of the journals *Circuits, Systems, and Signal Processing* and *The Journal of the Franklin Institute* and was an Associate Editor of the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING. He serves as the Chairman of the Connecticut IEEE Signal Processing chapter and was corecipient with P. Stoica of the 1989 IEEE Signal Processing Society's Senior Award for an outstanding paper. He is a member of Sigma Xi.

Technical Notes and Correspondence

Nonlinear L_1 Optimal Controllers for Linear Systems

A. A. Stoorvogel

Abstract—In this paper we study the L_1 optimal control problem for linear systems. We will show that by allowing the class of controllers to include nonlinear controllers, we can make the closed-loop L_1 norm strictly smaller than we could do using only linear controllers.

I. INTRODUCTION

The L_1 optimal control problem has been studied extensively in the literature. The L_1 problem was originally formulated in [1] and [8]. A solution to the problem was presented in [3] (for discrete-time systems) and [2] (for continuous-time systems). In these papers it became obvious that, when searching for optimal linear controllers for these problems, we need infinite-dimensional (continuous-time) or very high order (discrete-time) compensators. Especially for discrete-time systems, there is now a good theory available for the design of linear compensators (see e.g., [5]). The approach taken in these papers, however, is a method based on linear programming, and the method is therefore essentially constrained to linear compensators.

The objective of the current paper is to study whether we can improve by extending the class of compensators to include nonlinear compensators. For H_∞ , it is known, for instance, that this is not possible: the minimum over all stabilizing linear compensators of the closed-loop H_∞ norm is equal to the minimum over all stabilizing (possibly) nonlinear compensators (see [6]).

In [7] it has been shown that, although for the L_1 optimal control problem optimal and near optimal linear state feedback compensators are in general dynamic, there always exists static nonlinear compensators which achieve the same or better performance. Moreover, in [4] it was shown that nonlinear controllers which are differentiable in the origin cannot do better than linear controllers. Via an example, it was shown in that paper that nonlinear controller can do better for individual disturbances w . But, for this example, the worst-case L_1 norm could not be improved via nonlinear controllers. The objective of this paper is to show that we can achieve smaller L_1 norm if we allow for nonlinear, continuous controllers. This will be shown by means of a very simple static example. Hence the example applies equally well to continuous and discrete time. For ease of exposition we will concentrate on continuous-time systems only.

The paper has the following structure. In the next section we will give a problem formulation. Then we present our example and we conclude with some final remarks.

II. PROBLEM FORMULATION

We will consider systems of the form

$$\begin{cases} \dot{x} = Ax + Bu + Ew, \\ y = C_1x + D_{11}u + D_{12}w, \\ z = C_2x + D_{21}u + D_{22}w. \end{cases} \quad (2.1)$$

Manuscript received March 1, 1994.

The author is with the Department of Mathematics and Comp. Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

IEEE Log Number 9408781.

We will assume that x , u , w , y , and z take values in finite-dimensional vector spaces $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $w(t) \in \mathbb{R}^l$, $y(t) \in \mathbb{R}^q$, and $z(t) \in \mathbb{R}^p$. A special case are static systems where x is absent ($n = 0$) and we just have

$$\Sigma: \begin{cases} y = D_{11}u + D_{12}w, \\ z = D_{21}u + D_{22}w. \end{cases} \quad (2.2)$$

For a vector in \mathbb{R}^n we define the L_∞ -norm by

$$\|p\|_\infty = \sup |p_i| \quad (2.3)$$

where $p = (p_1, p_2, \dots, p_n)^T$. We define the function space L_∞ as the class of time-signals f for which the norm

$$\|f\|_\infty = \sup_{t \in \mathbb{R}^+} \|f(t)\|_\infty$$

is finite. This norm will be referred to as the L_∞ norm. We define the L_∞ -induced operator norm of an operator \mathcal{G} mapping w to z by

$$\|\mathcal{G}\|_1 = \sup_{w \neq 0, t \in \mathbb{R}^+} \frac{\|z\|_\infty}{\|w\|_\infty}$$

This norm is also referred to as the L_1 norm of \mathcal{G} . This is due to the fact that for operators \mathcal{G} described by a linear time invariant system of the form

$$\Sigma: \begin{cases} \dot{x} = Fx + Gw, \\ z = Hx + Ju \end{cases} \quad (2.4)$$

we find that the L_∞ -induced operator norm is equal to the L_1 norm of the impulse response which is defined by

$$\|H\|_1 = \int_0^\infty \|H(t)\|_\infty dt$$

where for a matrix $M = \{M_{ij}\}$ we have

$$\|M\| = \max_i \sum_j |M_{ij}|.$$

Clearly, this interpretation only holds for linear time-invariant systems and in particular does not hold for the closed-loop system we obtain by applying a nonlinear controller to (2.1). Nevertheless, since the L_∞ -induced operator norm yields a natural extension of the L_1 norm to nonlinear systems, we will often refer to the L_∞ -induced operator norm of a nonlinear system as the L_1 norm.

Note that for a static, time-invariant system like (2.2), first of all, it is obviously of no use to consider dynamic compensators. Hence we only consider static, time invariant but possibly nonlinear controllers. Then the output at time t of the closed-loop system is only affected by the disturbance at time t and the input at time t , and we obtain a finite dimensional optimization problem. Find a function K from \mathbb{R}^l to \mathbb{R}^m such that $I - D_{11}K$ is invertible and the closed-loop operator G_{cl} from \mathbb{R}^l to \mathbb{R}^p defined by

$$G_{cl} := D_{22} + D_{21}K(I - D_{11}K)^{-1}D_{12}$$

has minimal L_∞ -induced operator norm where the L_∞ norm of the input and output vectors are defined by (2.3)

III. EXAMPLE

In this section we will study the following very simple example

$$\begin{cases} y = u \\ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} u + \begin{bmatrix} 0 & 3 & 0 \\ -1.5 & 1.5 & -3 \\ -3 & 0 & 0 \\ 0 & 0 & -3 \end{bmatrix} \end{cases} \quad (3.1)$$

We claim that for this static example the minimal achievable I_1 norm is 3. The latter can only be achieved, however, via a nonlinear controller. Via a linear controller the minimal achievable I_1 norm is equal to 3.75.

A. Linear Controllers

We will study the optimal input u for four specific disturbances

$$\begin{aligned} u_1 &= \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, & u_2 &= \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \end{pmatrix} \\ u_3 &= \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, & u_4 &= \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \end{aligned} \quad (3.2)$$

Each of these disturbances has norm one. Hence to achieve an I_1 norm equal to 3 we have to have that for u and the corresponding control input u we obtain an output with norm less than or equal to 3. For u_1 , u_2 and u_3 we find that $\|y\|_\infty = 3 + \|u\|_\infty$. For u_4 we have $\|y\|_\infty = \max\{6 - u_1, u_1 - 3\}$. Since we only allow linear controllers and $u_1 = u_1 - u = -u_4$, we must have $u_1 = u_1 - u = -u_4$. We find that to make $\|y\|_\infty < 3.75$ then either $u_1 = u$ or u_4 must be larger than 0.75 and hence u_1 or u_4 must have norm larger than 3.75. This implies that we can never make the I_1 norm less than 3.75 by linear controllers. Moreover, the linear controller

yields the closed loop system

$$\begin{bmatrix} 0.75 & 2.25 & 0.75 \\ -0.75 & 0.75 & -2.25 \\ -2.25 & -0.75 & 0.75 \\ 0.75 & -0.75 & -2.25 \end{bmatrix}$$

and it clearly achieves a closed loop I_1 norm of 3.75.

B. Nonlinear Controllers

To find an optimal nonlinear controller for the system (3.1) we again look at the four disturbances from (3.2). Since we have $\|y\|_\infty = 3 + \|u\|_\infty$, it is obvious that we cannot make the I_∞ induced operator norm less than three. We will now construct an optimal nonlinear controller which achieves an I_1 norm equal to three. Consider the eight corners of the cube $\|u\|_\infty = 1$. Besides the four corners given by (3.2) we have

$$\begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \quad u_5 = \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$$

$$u_6 = \begin{bmatrix} 1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \quad u_7 = \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{bmatrix} \quad (3.3)$$

We find that the optimal input for each corner is $u_1 = 0, u_2 = 0, u_3 = 0, u_4 = 3, u_5 = 0, u_6 = 0, u_7 = 0, u_8 = -3$. This yields an output with norm three for each corner. We next extend our function to each face of the cube. For any u with $\|u\|_\infty = 1$ we find

$$u = \lambda_1 u_1 + \lambda_2 u_2 + \lambda_3 u_3 + \lambda_4 u_4$$

for some $\lambda_i \geq 0$ with $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$. The choice of λ_i is not unique but it is straightforward to derive a continuous selection rule. We choose the corresponding input as

$$u = \lambda_1 u_1 + \lambda_2 u_2 + \lambda_3 u_3 + \lambda_4 u_4$$

and the resulting output is

$$y = \lambda_1 y_1 + \lambda_2 y_2 + \lambda_3 y_3 + \lambda_4 y_4$$

Finally we note that

$$\begin{aligned} \|y\|_\infty &\leq \lambda_1 \|y_1\|_\infty + \lambda_2 \|y_2\|_\infty + \lambda_3 \|y_3\|_\infty + \lambda_4 \|y_4\|_\infty \\ &= 3 + 3\|u\|_\infty \end{aligned}$$

In this way we have defined a continuous function f on the cube $\|u\|_\infty = 1$. We can extend this function to the whole \mathbb{R}^4 by

and the closed loop system resulting from $u = f(u)$ can be easily checked to have I_1 induced operator norm equal to three. We have already seen that this feedback is therefore optimal.

IV. CONCLUSION

In previous papers the first step was to use the Youla parameterization to bring the closed loop operator in the form $T_1 + T_2 Q I_1$, where Q is the design variable which should be a stable system and which determines uniquely the corresponding controller. In [4] it has been shown that for $I_1 = I$ nonlinear controllers can not achieve a smaller I_1 norm. This paper gives an example where $T_1 = I$ and we can do better by nonlinear controllers. $T_1 = I$ is connected to estimation problems while $T_2 = I$ is connected to control problems. This shows a clear lack of duality and the best we can hope for is a kind of separation structure for I_1 controllers which will be of the form of a linear estimator interconnected to a nonlinear static state feedback. Nonlinear static state feedbacks were already studied in [7].

This paper basically only tells us that we should study nonlinear compensators if we want to obtain optimal controllers.

REFERENCES

- [1] A. I. Barabanov and O. N. Granichin, 'Optimal controller for a linear plant with bounded noise', *Avtomatika i Telemekhanika*, vol. 5, pp. 39-46, 1984.
- [2] M. A. Dahleh and J. B. Pearson, ' H^1 optimal compensators for continuous time systems', *IEEE Trans Automat Contr*, vol. 32, pp. 889-895, 1987.
- [3] ———, ' H^1 optimal compensators for MIMO discrete time systems', *IEEE Trans Automat Contr*, vol. 32, pp. 314-322, 1987.
- [4] M. A. Dahleh and J. S. Shamma, 'Rejection of persistent bounded disturbances: Nonlinear controllers', *Syst Contr Lett*, vol. 18, pp. 245-252, 1992.

- [5] I. J. Diaz-Bobillo and M. A. Dahleh, "Minimization of the maximum peak-to-peak gain: the general multiblock problem," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 1459–1482, 1993.
- [6] P. P. Khargonekar and K. R. Poolla, "Uniformly optimal control of linear time-varying plants: nonlinear time-varying controllers," *Syst. Contr. Lett.*, vol. 5 pp. 303–308, 1986.
- [7] J. S. Shamma, "Nonlinear state feedback for ℓ^1 optimal control," *Syst. Contr. Lett.*, vol. 21, pp. 265–270, 1993.
- [8] M. Vidyasagar, "Optimal rejection of persistent bounded disturbances," *IEEE Trans. Automat. Contr.*, vol. 31, pp. 527–534, 1986.

Simultaneous Observation of Linear Systems

Y. X. Yao, M. Darouach, and J. Schaefer

Abstract—This note considers the problem of designing an observer which can observe the states for each of a given set of plants. The concept of the simultaneous observation is introduced. The simultaneous observation is achieved using coprime factorization techniques. Necessary and sufficient conditions for the existence of simultaneous observation are derived. The observer design is discussed, and an example is given.

I. INTRODUCTION

In this note we are interested in the problem of designing an observer which can observe the states for each of a given set of plants. Specifically, suppose that $G_0(s), G_1(s), \dots, G_n(s)$ are a finite number of given plants, we would like to know whether or not there exists a common observer for this given set of plants. This is referred to as the simultaneous observation problem in this note.

There are two reasons at least to study the simultaneous observation problem like the simultaneous stabilization problem [1] in which a compensator stabilizes each of a given set of plants. First, because the structure of the plant is often subjected to unexpected change, such as components fault or variation in operating condition, we wish to design an observer which can continue to observe the states of the plant in the presence of the plant change. Specifically, $G_0(s)$ corresponds to the nominal plant description to be observed, while $G_1(s), G_2(s), \dots, G_n(s)$ correspond to the plant description after some sort of structural change. The problem to design an observer which can continue to observe the states of the plant after this sort of structural change is equivalent to find an observer which can observe the states for each of the plants $G_0(s), G_1(s), \dots, G_n(s)$. Secondly, it may arise in the observer design of a nonlinear system at various operating conditions. The nonlinear system is often linearized at the various operating points, and some of linearized models can be obtained. If the states of each of these linearized models can be observed using a common observer, then a fixed observer can be used for the nonlinear system.

Manuscript received November 30, 1993; revised May 17, 1994 and August 2, 1994.

Y. X. Yao is with the Centre de Recherche Public Henri Tudor, 6, rue Coudenhove-Kalergi, L-1359, Luxembourg and the EARAL CRAN CNRS UA-821, 186 rue de Lorraine, 54400 Cosnes-et-Romain, France.

M. Darouach is with the EARAL CRAN CNRS UA-821, 186 rue de Lorraine, 54400 Cosnes-et-Romain, France.

J. Schaefer is with the Centre de Recherche Public Henri Tudor, 6, rue Coudenhove-Kalergi, L-1359, Luxembourg.

IEEE Log Number 9408777.

The proposed problem is motivated by the simultaneous control system design problem, particularly the simultaneous stabilization problem which was first posed in [2] and [3]. For a given set of plants, one would like to know whether or not there exists a common controller such that all plants are simultaneously stabilized in the simultaneous stabilization problem. It is also natural that one would like to know whether or not there exists a common observer such that the states for each of given plants are simultaneously observed.

This note focuses on the existence of simultaneous observation. The observer design is also discussed.

II. PROBLEM DESCRIPTION AND PRELIMINARIES

Consider a set of the linear time-invariant systems described by

$$\dot{x}(t) = A_i x(t) + B_i u(t) \quad (1)$$

$$y(t) = C_i x(t) + D_i u(t), \quad i = 0, 1, \dots, n \quad (2)$$

where $x(t) \in R^n$ is the state vector, $u(t) \in R^p$ is the input vector and $y \in R^m$ is the observation vector, and A_i, B_i, C_i and D_i are constant matrices of the appropriate dimensions. Taking Laplace transformation of (1) and (2), gives

with

$$G_i(s) = C_i(sI - A_i)^{-1} B_i + D_i.$$

It is desired to design an observer which will produce an estimation for

$$\hat{x}(t) = L_i x(t) \quad (4)$$

with $L_i \in R^{k \times n}$. An observer for the systems (1)–(3) is expressed as a dynamic system

$$\dot{r}(s) = F(s)u(s) + H(s)y(s) \quad (5)$$

with the property that the estimation error satisfies

$$\lim_{t \rightarrow \infty} (L_i x(t) - r(t)) = 0 \quad (6)$$

for all $u(s)$, where $F(s)$ and $H(s)$ are RH_∞ matrices, i.e., they are stable transfer function matrices. If such an observer exists, we say that $G_0(s), G_1(s), \dots, G_n(s)$ in (3) are simultaneously observable.

Our main problem to be solved is: given $n+1$ plants as (1), (2), or (3), we would like to know whether or not there exists a common observer for this given set of plants. This is referred to as the simultaneous observation problem.

The right-coprime and left-coprime factorization of the i th plant G_i are

$$G_i(s) = N_i(s)M_i^{-1}(s) = \hat{M}_i^{-1}(s)\hat{N}_i(s) \quad (7)$$

respectively, and $N_i(s), M_i(s)$ are right-coprime RH_∞ matrices, $\hat{M}_i(s)$ and $\hat{N}_i(s)$ are left-coprime RH_∞ matrices. For the double-coprime factorization of $G_i(s)$ there exist RH_∞ transfer functions $Y_i(s), X_i(s), \hat{Y}_i(s)$ and $\hat{X}_i(s)$ such that

$$\begin{aligned} & \begin{bmatrix} Y_i(s) & X_i(s) \\ -\hat{N}_i(s) & \hat{M}_i(s) \end{bmatrix} \begin{bmatrix} M_i(s) & -\hat{X}_i(s) \\ N_i(s) & Y_i(s) \end{bmatrix} \\ &= \begin{bmatrix} M_i(s) & \hat{X}_i(s) \\ -N_i(s) & \hat{Y}_i(s) \end{bmatrix} \begin{bmatrix} Y_i(s) & -X_i(s) \\ \hat{N}_i(s) & \hat{M}_i(s) \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \end{aligned} \quad (8)$$

Each matrix in Bezout identity (8) can be calculated by using the existing algorithms given in [1]. The state-space construction of the right-coprime factors $N_i(s)$ and $M_i(s)$ is given in the following

$$N_i(s) = C_{ki}(sI - A_{ki})^{-1}B_i + D_i, \quad (9)$$

$$M_i(s) = K_i(sI - A_{ki})^{-1}B_i + J \quad (10)$$

where K_i is matrix such that $A_{ki} = A_i + B_i K_i$ is stable and $C_{ki} = C_i + D_i K_i$.

Define the partial-state $\xi(s)$ as follows

$$M_i(s)\xi(s) = u(s).$$

Then from (7), (3) can be written as

$$M_i(s)\xi(s) = u(s) \quad (11)$$

$$y(s) = N_i(s)\xi(s) \quad (12)$$

and the variable $E_i x(s)$ according to (4) can be expressed by

$$z(s) = P_i(s)\xi(s) \quad (13)$$

where $P_i(s)$ can be calculated as follows.

From (7), the right-coprime factorization of the i th plant G_i can be written as

$$C_i(sI - A_i)^{-1}B_i + D_i = N_i(s)M_i^{-1}(s).$$

Using (9) and (10) we have

$$(sI - A_i)^{-1}B_i = (sI - A_{ki})^{-1}B_i M_i^{-1}(s) \quad (14)$$

which leads to

$$x(s) = (sI - A_i)^{-1}B_i u(s) = (sI - A_{ki})^{-1}B_i \xi(s).$$

Thus, the variable $z(s)$ in (13) is obtained with RH_∞ matrices

$$P_i(s) = E_i(sI - A_{ki})^{-1}B_i. \quad (15)$$

Then the problem of designing a common observer for the systems (1), (2), or (3) can be stated as finding RH_∞ -matrices $F(s)$ and $H(s)$ such that for all $u(s)$

$$z(s) - r(s) = 0. \quad (16)$$

Condition (16) is satisfied, i.e., $E_i x(t)$ can be observed if and only if the following condition holds [5]

$$F(s)M_i(s) + H(s)N_i(s) = P_i(s). \quad (17)$$

When only one plant is considered in (1), (2), or (3), the parameterization of the observer (5) has been given in [4], [6] using the factorization approach [1]. The result is given in the following lemma.

Lemma 1 [6]: The set of RH_∞ matrices $F(s)$, $H(s)$ that satisfy

$$F(s)M(s) + H(s)N(s) = P(s) \quad (18)$$

is given by

$$F(s) = P(s)Y(s) - Q(s)\hat{N}(s), \quad (19)$$

$$H(s) = P(s)X(s) + Q(s)\hat{M}(s), \quad Q(s) \in RH_\infty \quad (20)$$

and, for every RH_∞ matrix $Q(s)$ of appropriate dimensions, $F(s)$ and $H(s)$ satisfying (19) and (20) also fulfill the existence condition (18) of the observer.

Based on this parameterization, we now mainly consider the problem for the existence of the simultaneous observation. We would like to know whether or not there exists a common observer with the form (5) satisfying the observer existence condition (17) for the given set of plants (1), (2), or (3), and if this observer exists, how to design it?

III. MAIN RESULTS

First, we study the problem of simultaneously observing the states of two plants which can be stated as follows: Given two plants $G_0(s)$ and $G_1(s)$, when does there exist a common observer such that they are both observable?

Theorem 1: Given two plants $G_0(s)$ and $G_1(s)$, define

$$A(s) = Y_0(s)M_1(s) + X_0(s)N_1(s), \quad (21)$$

$$B(s) = -\hat{N}_0(s)M_1(s) + \hat{M}_0(s)N_1(s). \quad (22)$$

Then $G_0(s)$ and $G_1(s)$ can be simultaneously observed if and only if there exists an $R(s) \in RH_\infty$ such that

$$P_0(s)A(s) + R(s)B(s) = P_1(s) \quad (23)$$

holds.

Proof: From Lemma 1, the sets of observers that can observe $G_0(s)$ and $G_1(s)$ are given by

$$\begin{aligned} r(G_0) &= \{F_0(s) = P_0(s)Y_0(s) - Q_0(s)\hat{N}_0(s), H_0(s) \\ &= P_0(s)X_0(s) + Q_0(s)\hat{M}_0(s); Q_0(s) \in RH_\infty\}, \end{aligned} \quad (24a)$$

$$\begin{aligned} r(G_1) &= \{F_1(s) = P_1(s)Y_1(s) - Q_1(s)\hat{N}_1(s), H_1(s) \\ &= P_1(s)X_1(s) + Q_1(s)\hat{M}_1(s); Q_1(s) \in RH_\infty\}. \end{aligned} \quad (24b)$$

Hence $G_0(s)$ and $G_1(s)$ can be simultaneously observed if and only if there exist $Q_0(s)$, $Q_1(s) \in RH_\infty$ such that

$$P_0(s)Y_0(s) - Q_0(s)\hat{N}_0(s) = P_1(s)Y_1(s) - Q_1(s)\hat{N}_1(s), \quad (25)$$

$$P_0(s)X_0(s) + Q_0(s)\hat{M}_0(s) = P_1(s)X_1(s) + Q_1(s)\hat{M}_1(s). \quad (26)$$

Now (25), (26) can be rewritten as

$$\begin{aligned} [P_0(s) \quad Q_0(s)] \begin{bmatrix} Y_0(s) & X_0(s) \\ -\hat{N}_0(s) & \hat{M}_0(s) \end{bmatrix} \\ = [P_1(s) \quad Q_1(s)] \begin{bmatrix} Y_1(s) & X_1(s) \\ -\hat{N}_1(s) & \hat{M}_1(s) \end{bmatrix} \end{aligned} \quad (27)$$

which is equivalent to

$$\begin{aligned} [P_0(s) \quad Q_0(s)] \begin{bmatrix} Y_0(s) & X_0(s) & M_1(s) & -\hat{N}_1(s) \\ -\hat{N}_0(s) & \hat{M}_0(s) & N_1(s) & \hat{Y}_1(s) \end{bmatrix} \\ = [P_1(s) \quad Q_1(s)] \end{aligned} \quad (28)$$

from (8), or

$$[P_0(s) \quad Q_0(s)] \begin{bmatrix} A(s) & S(s) \\ B(s) & T(s) \end{bmatrix} = [P_1(s) \quad Q_1(s)] \quad (29)$$

where

$$\begin{bmatrix} A(s) & S(s) & Y_0(s) & X_0(s) \\ B(s) & T(s) & -\hat{N}_0(s) & \hat{M}_0(s) \end{bmatrix} \begin{bmatrix} M_1(s) & -\hat{N}_1(s) \\ N_1(s) & \hat{Y}_1(s) \end{bmatrix} \quad (30)$$

Thus $G_0(s)$ and $G_1(s)$ can be simultaneously observed if and only if there exists $Q_0(s)$, $Q_1(s) \in RH_\infty$ such that (29) holds. Therefore Theorem 1 is proved if we can establish that (29) holds if and only if $P_0(s)A(s) + R(s)B(s) = P_1(s)$ for some $R(s) \in RH_\infty$.

Suppose (29) holds for suitable $Q_0(s)$, $Q_1(s) \in RH_\infty$, then $P_0(s)A(s) + Q_0(s)B(s) = P_1(s)$. Conversely, suppose $P_0(s)A(s) + R(s)B(s) = P_1(s)$ for some $R(s) \in RH_\infty$, then (29) holds with $Q_0(s) = R(s)$, $Q_1(s) = (S(s) + Q_0(s)T(s))$. This completes the proof. \square

Theorem 1 shows that the problem of simultaneously observing two plants can be reduced to that of observing a single plant using a common observer.

It should be pointed out that the proof of Theorem 1 closely follows the corresponding proof of the simultaneous stabilization problem for two plants (e.g., see [1, Theorem 5.4.2]). In addition, it is seen that $A(s)$ and $B(s)$ in Theorem 1 are the same as those of Theorem

5.4.2 in [1]. It is shown in [1] that $A(s)$ and $B(s)$ are right coprime. Theorem 5.4.2 in [1] states that two systems $G_0(s)$ and $G_1(s)$ can be simultaneously stabilized if and only if the associated system $\hat{G}(s) = B(s)A(s)^{-1}$ is strongly stabilizable, i.e., it can be stabilized using a stable compensator.

Corollary 1: If condition (23) is satisfied for two given plants $G_0(s)$ and $G_1(s)$ in Theorem 1, then

$$F_0(s)M_1(s) + H_0(s)N_1(s) = P_1(s). \quad (31)$$

Proof: This is a direct result from Theorem 1; its proof is omitted here. \square

This corollary further indicates that the observer of the first plant $G_0(s)$ can be used to observe the second plant $G_1(s)$. Therefore two plants can be observed by a common observer. It is a more intuitive result in the following.

Corollary 2: Suppose $G_0(s)$ is stable and $G_1(s)$ is arbitrary. Then $G_0(s)$ and $G_1(s)$ can be simultaneously observed if and only if $G_1(s) - G_0(s)$ can be observed.

Proof: Suppose $\hat{M}_0^{-1}(s)\hat{N}_0(s)$ is a left-coprime factorization of $G_0(s)$ and $N_1(s)M_1^{-1}(s)$ is a right-coprime factorization of $G_1(s)$. If $G_0(s)$ is stable, then $\hat{N}_0(s) = G_0(s)$, $\hat{M}_0(s) = I$, $Y_0(s) = I$ and $X_0(s) = 0$. From

$$\begin{aligned} \hat{G}(s) &= B(s)A(s)^{-1} \\ &= (N_1(s) - G_0(s)M_1(s))M_1(s)^{-1} \\ &= G_1(s) - G_0(s) \end{aligned} \quad (32)$$

it can be seen that $(N_1(s) - G_0(s)M_1(s))M_1(s)^{-1}$ is a right-coprime factorization of $G_1(s) - G_0(s)$. From Theorem 1 it follows that

$$P_0(s)M_1(s) + R(s)(N_1(s) - G_0(s)M_1(s)) = P_1(s) \quad (33)$$

which can be written as

$$(P_0(s) - R(s)G_0(s))M_1(s) + R(s)N_1(s) = P_1(s) \quad (34)$$

define

$$\hat{F}(s) = (P_0(s) - R(s)G_0(s)) \quad (35)$$

$$\hat{H}(s) = R(s) \quad (36)$$

then (34) shows that $r(s) = \hat{F}(s)u(s) + \hat{H}(s)y(s)$ can observe the plant $G_1(s)$. It also can be proved that

$$(P_0(s) - R(s)G_0(s))M_0(s) + R(s)N_0(s) = P_0(s). \quad (37)$$

That is $r(s) = \hat{F}(s)u(s) + \hat{H}(s)y(s)$ can also observe the plant $G_0(s)$. Hence $G_0(s)$ and $G_1(s)$ can be simultaneously observed. \square

It can be seen that from (19) and (20), in fact, $r(s) = \hat{F}(s)u(s) + \hat{H}(s)y(s)$ is the observer of the first plant $G_0(s)$ with $Q_0(s) = R(s)$.

Next we turn to the simultaneous observation problem with more than two plants. We would like to know whether or not there exists an observer that can observe the states of each one of a given set of plants $G_0(s), G_1(s), \dots, G_n(s)$. In the following theorem, we give necessary and sufficient condition to observe simultaneously $n + 1$ plants using a common observer.

Theorem 2: Suppose G_0, G_1, \dots, G_n are given plants, define

$$A_i(s) = Y_0(s)M_i(s) + X_0(s)N_i(s), \quad (38)$$

$$B_i(s) = -\hat{N}_0(s)\hat{M}_i(s) + \hat{M}_0(s)N_i(s), \quad i = 1, 2, \dots, n. \quad (39)$$

Then $G_0(s), G_1(s), \dots, G_n(s)$ can be simultaneously observed if and only if there exists a $R(s) \in RH_\infty$, such that

$$P_0(s)A_i(s) + R(s)B_i(s) = P_i(s) \quad (40)$$

holds for $i = 1, 2, \dots, n$.

Proof: The proof closely parallels that of Theorem 1; it is omitted here. \square

It is shown in Theorem 2 that the problem of simultaneously observing $n + 1$ plants is equivalent to that of simultaneously observing n plants with a common observer.

Now suppose $R(s) \in RH_\infty$ exists which can observe the auxiliary plant $\hat{G}(s)$, then from (5), (19), and (20), the common observer for the plants $G_0(s)$ and $G_1(s)$ is just

$$\begin{aligned} r(s) &= [P_0(s)Y_0(s) - R(s)\hat{N}_0(s)]u(s) \\ &\quad + [P_0(s)X_0(s) + R(s)\hat{M}_0(s)]y(s). \end{aligned} \quad (41)$$

In the assumption of Corollary 2, the plant $G_0(s)$ is stable, then from (35) and (36), the observer has the following form

$$r(s) = (P_0(s) - R(s)G_0(s))u(s) + R(s)y(s), \quad R(s) \in RH_\infty \quad (42)$$

where $R(s)$ can be obtained from (23).

IV. DESIGN EXAMPLE

Consider the system

$$\begin{bmatrix} 0 & 1 \\ a_{21} & a_{22} \end{bmatrix} x(t) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t) \quad (43)$$

$$y = [2 \quad 7]x(t) \quad (44)$$

$$z = [1 \quad -1]x(t). \quad (45)$$

The parameter values (a_{21}, a_{22}) for the two operating conditions are.

$$1) (a_{21}, a_{22}) = (-1, -2);$$

$$2) (a_{21}, a_{22}) = (-2, -3).$$

For these conditions, we get

$$G_0(s) = \frac{2s-3}{s^2+2s+1}, \quad G_1(s) = \frac{2s-8}{s^2+3s+2}. \quad (46)$$

Because the plant $G_0(s)$ is stable, the related elements of double coprime factorization for the first plant are

$$\hat{N}_0(s) = \frac{2s-3}{s^2+2s+1} = G_0(s),$$

$$M_0(s) = 1, \quad X_0(s) = 0, \quad Y_0(s) = 1,$$

$$P_0(s) = \frac{s+3}{s^2+2s+1}$$

and the related elements of double-coprime factorization for the second plant are

$$M_1(s) = \frac{s^2+3s+2}{s^2+2s+3}, \quad N_1(s) = \frac{2s-8}{s^2+2s+3},$$

$$P_1(s) = \frac{s+5}{s^2+2s+3}.$$

From (21) and (22) in Theorem 1, we have

$$A(s) = Y_0(s)M_1(s) + X_0(s)N_1(s) = \frac{s^2+3s+2}{s^2+2s+3}, \quad (47)$$

$$\begin{aligned} B(s) &= -\hat{N}_0(s)M_1(s) + \hat{M}_0(s)N_1(s) \\ &= \frac{-(7s^2+9s+2)}{(s^2+2s+3)(s^2+2s+1)}. \end{aligned} \quad (48)$$

From (23), we get

$$R(s) = \frac{-s+1}{7s+2} \in RH_\infty. \quad (49)$$

Then from (42), the observer is given by

$$r(s) = \frac{9}{7s+2}u(s) + \frac{-s+1}{7s+2}y(s). \quad (50)$$

The observer (50) can observe simultaneously the two plants in (46).

V. CONCLUSION

In this note, a class of new observation problem, called simultaneous observation problem, is introduced and studied. This is achieved by using the coprime factorization approach. Necessary and sufficient conditions for the existence of simultaneous observation are derived. It was shown that a general solution for the simultaneous observation problem reduces the problem of observing n plants using a common observer to one of observing $n - 1$ plants using a common observer.

REFERENCES

- [1] M. Vidyasagar, *Control Systems Synthesis: A Factorization Approach*, Cambridge, MA: MIT Press, 1985.
- [2] P. Sacks and J. Murray, "Fractional representation, algebraic geometry and the simultaneous stabilization problem," *IEEE Trans Automat Contr.*, vol. AC-27, pp. 895-903, 1982.
- [3] M. Vidyasagar and N. Viswanadham, "Algebraic design techniques for reliable stabilization," *IEEE Trans Automat Contr.*, vol. AC-27, pp. 1085-1095, 1982.
- [4] X. Ding, P. M. Frank, and L. Guo, "Robust observer design for dynamical systems under unknown disturbances," in *Proc. First IFAC Symp. Design Methods Contr. Syst.*, ETH Zurich, Switzerland, 1991, pp. 290-295.
- [5] J. O'Reilly, *Observer for Linear Systems*, London: Academic, 1983.
- [6] X. Ding and P. M. Frank, "Fault detection via factorization approach," *Syst. Cont. Lett.*, vol. 14, pp. 431-436, 1990.

An Algorithm for Computing the Mask Value of the Supremal Normal Sublanguage of a Legal Language

Michel Barbeau, Guy Casteau, and Richard St-Denis

Abstract—We consider the problem of finding the mask value of the supremal normal sublanguage L_N of some given language L . We describe a straightforward algorithmic solution that can be applied to existing off-line procedures for determining the supremal controllable and normal sublanguage of L and that does not require an explicit calculation of L_N . This problem is fundamental because it is related to the supervisory control problem under partial observation. Our algorithm applies only to closed languages.

I. INTRODUCTION

One of the basic problems in supervisory control is to design a controller whose task is to enable and disable the controllable events of a discrete-event system (DES) such that the control requirements expressed as a legal language L are satisfied. A unifying theory has been developed by Ramadge and Wonham [11] to define and solve this problem. Lin and Wonham [10] and Cieslak *et al.* [6] have studied, in the framework of Ramadge and Wonham, the supervisory control problem under partial observation. Following the theory of Ramadge and Wonham, the uncontrolled DES is represented by a generator, which is a deterministic automaton $G = (Q, \Sigma, \delta, q_0, Q_m)$, where Q is a set of states, Σ a finite set of events, $\delta: \Sigma \times Q \rightarrow Q$

a transition function, $q_0 \in Q$ an initial state, and $Q_m \subseteq Q$ a set of marked states. Let $\Gamma = \{0, 1\}^\Sigma$ be the set of all binary assignments to the elements of Σ , and let $\gamma \in \Gamma$ and $\sigma \in \Sigma$. If $\gamma(\sigma) = 1$ then σ is enabled; otherwise, σ is disabled. Let Σ_c and Σ_u be fixed disjoint subsets of Σ denoting the sets of controllable and uncontrollable events respectively, such that $\Sigma = \Sigma_c \cup \Sigma_u$. A controller is a pair $C = (S, \phi)$, where $S = (Y, \Lambda, \zeta, y_0, Y_m)$ is a deterministic automaton and $\phi: Y \rightarrow \Gamma$ is a feedback function satisfying: i) $\phi(y)(\sigma) = 1$, if $\sigma \in \Sigma_u$; and ii) $\phi(y)(\sigma) \in \{0, 1\}$, otherwise.

To control a DES, Ramadge and Wonham introduced a transition function $\delta_c: \Gamma \times \Sigma \times Q \rightarrow Q$ defined by

$$\delta_c(\gamma, \sigma, q) = \begin{cases} \delta(\sigma, q), & \text{if } \delta(\sigma, q) \text{ is defined and } \gamma(\sigma) = 1 \\ \text{undefined}, & \text{if } \delta(\sigma, q) \text{ is undefined or } \gamma(\sigma) = 0. \end{cases}$$

The controlled discrete-event system (CDES) is then the generator $G_c = (Q, \Gamma \times \Sigma, \delta_c, q_0, Q_m)$ obtained from G by specifying the sets Σ_c and Σ_u .

Following the extensions proposed by Cieslak *et al.* [6], we consider the case in which C observes all the events of G , through a mask or observation function M that maps each event in Σ into an observed event in $\Lambda \cup \{-\}$. The events in $M^{-1}(\cdot)$ are those that cannot be seen by C . If C cannot distinguish between σ_1 and σ_2 , then $M(\sigma_1) = M(\sigma_2)$. If M is the identity function, then $\Lambda = \Sigma$ and all the original events are observed by C . The special case in which M simply erases some of the events in Σ occurs frequently and is called a natural projection [12] or natural mask [14].

Finally, the CDES and controller are embodied in a closed-loop system to constitute a supervised discrete-event system (SDES) C/G_c , which is defined to be the generator $(Y \times Q, \Sigma, (\zeta \circ M) \times \delta_c, (y_0, q_0), Y_m \times Q_m)$, where the function

$$(\zeta \circ M) \times \delta_c: \Sigma \times Y \times Q \rightarrow Y \times Q$$

is defined as $((\zeta \circ M) \times \delta_c)(\sigma, y, q) = (\zeta(M(\sigma), y), \delta_c(\phi(y), \sigma, q))$, if $\delta(\sigma, q)$ and $\zeta(M(\sigma), y)$ are defined, and $\phi(y)(\sigma) = 1$, and is undefined, otherwise.

Let $K \subseteq L \subseteq L(G) \subseteq \Sigma^*$ and $\Sigma' \subseteq \Sigma$. We recall that a language K is closed if $K = \bar{K}$, the prefix closure of K . In this paper, we assume that all languages are closed and denote by \bar{s} the prefix closure of a string $s \in \Sigma^*$. Language K is $(\Sigma', L(G))$ -controllable if $(\forall s \in K)(\forall \sigma \in \Sigma')[s\sigma \in L(G) \Rightarrow s\sigma \in K]$. Language K is $(M, L(G))$ -normal if $(\forall s \in K)(\forall s' \in L(G))[M(s) = M(s') \Rightarrow s' \in K]$. Finally, K is $(\Sigma', L(G))$ -observable if $(\forall s, s' \in K)(\forall \sigma \in \Sigma')[M(s) = M(s') \wedge s\sigma \in K \wedge s'\sigma \in L(G) \Rightarrow s'\sigma \in K]$.

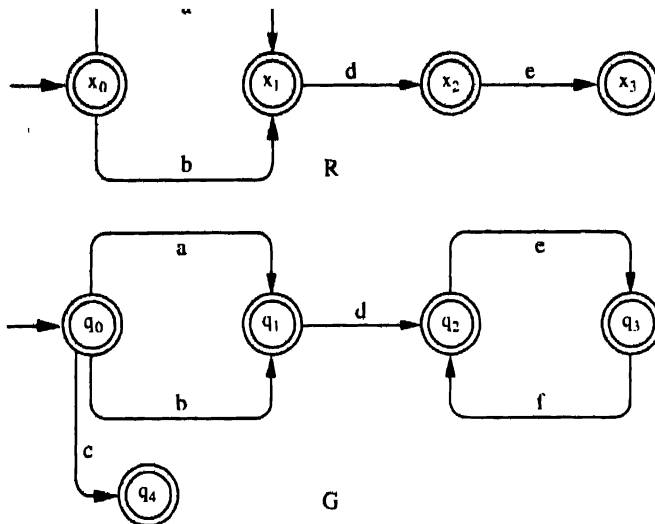
The supervisory control problem under partial observation is formally stated as follows. Given a CDES G_c , its legal behavior $L \subseteq L(G)$, and a mask function M , find a controller $C = (S, \phi)$ such that $L(C/G_c)$ is the largest sublanguage of L that is closed, $(\Sigma_u, L(G))$ -controllable and $(M, L(G))$ -normal. This problem is slightly different from the supervisory control and observation problem (SCOP) first formulated by Lin and Wonham [10], in which M is a natural mask and $L(C/G_c)$ is constrained to contain a minimal acceptable language.

Let us consider a simple example. Fig. 1 depicts generator G of the language of a plant and automaton R of its legal language L . The observation function is shown in Table I. Note that the illegal event c of the plant is observed as the legal event b . Note also that the words $adcf$ and $bdcf$ of the plant are observed as the legal words

Manuscript received December 20, 1993; revised August 1, 1994. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (FCAR).

The authors are with the Département de mathématiques et d'informatique, Université de Sherbrooke, Sherbrooke, Québec, Canada J1K 2R1.

IEEE Log Number 9408778.

Fig. 1. Automata R and G .

ade and bde , respectively, but that $adef$ and $bdef$ are not admitted. The largest normal sublanguage of L clearly is: $L_R = \{a, ad\}$. To make R an automaton for L_R , the transition from x_0 to x_1 on event b must be removed as well as state x_3 with the incoming transition from x_2 . In this way, indistinguishable behaviors of the plant are all permitted by the controller if they are all legal, or all rejected if at least one of them is illegal.

Cieslak *et al.* [6] proved, under the assumption that $M^{-1}(M(\Sigma_n) - \{\varepsilon\}) \subseteq \Sigma_n$, that K_R , the supremal controllable and normal language contained in L , can be computed as follows:

- Step 1) Compute L_R , the largest $(M, L(G))$ -normal sublanguage of L .
- Step 2) Compute $M(L_R)$, $M(L(G))$, and $\Delta_n = M(\Sigma_n) - \{\varepsilon\}$.
- Step 3) Compute K , the largest $(\Delta_n, M(L(G)))$ -controllable sublanguage of $M(L_R)$, by using the algorithm developed by Ramadge and Wonham [15].
- Step 4) Compute $K_R = M^{-1}(K) \cap L(G)$.

This paper presents an alternate procedure for computing the supremal controllable and normal sublanguage K_R , in which Step 1) above is omitted. That is to say, our procedure does not require an intermediate representation of L_R . It includes a new algorithm for computing directly the mask value $M(L_R)$ of the supremal normal sublanguage L_R . Indeed, $L(G)$, $M(L_R)$, and $M(L(G))$ are the only languages required for completing the last two steps of the procedure. Compared with the approach of Cieslak *et al.* [6], our procedure provides a simpler and more straightforward algorithmic solution.

This paper is organized as follows. Section II presents a review of known related algorithms. Section III gives a description of our algorithm and a proof of its correctness. Section IV provides a comparison with existing algorithms and contains concluding remarks.

II. RELATED WORK

The supervisory control problem under partial observation has received much attention, and several algorithms for computing the supremal controllable and normal sublanguage of a given language have been proposed in the literature [4]–[6], [14]. They can all be used for the off-line derivation of a controller, but the worst-case computational complexity of all these algorithms is theoretically exponential because there is no polynomial-time algorithm for the SCOP unless $P = NP$ [13]. The exponential-time complexity

TABLE I
THE OBSERVATION FUNCTION M

σ	a	b	c	d	e	f
$M(\sigma)$	a	b	b	e	e	e

results from the construction of one or more deterministic automata from nondeterministic automata. But, it should be noted that in practice the worst-case occurs rarely [1]. An on-line approach was, however, suggested by Heymann and Lin [7] for improving this bound. They consider the required control action to be calculated at each step of the actual execution of the closed-loop system. The basic idea behind their algorithm is that a controller that has been designed for operation under full observation can be modified to operate under partial observation. The computational complexity at each step is polynomial in the product of the number of states in automaton R for the legal language L and the number of states in G . Furthermore, the closed-loop behavior thus achieved is a controllable and observable sublanguage larger than the supremal controllable and normal sublanguage. It should be noted, however, that the normality and observability properties are equivalent under the assumptions that the language is controllable and all controllable events are observable [9].

Nevertheless, the off-line approach is useful when there is a need to design a full controller because the plant is highly time-critical or the closed-loop system must be verified prior to its execution. In addition to the off-line procedure of Cieslak *et al.* [6], Brandt *et al.* [4], [14] propose the algebraic formula $L - M^{-1}(M(L(G) - L))$ for the off-line computation of L_R . These operations can be interpreted as follows. Take the set of illegal words, that is, $L(G) - L$. Compute how the illegal words are observed, that is, $M(L(G) - L)$. Compute the set of words, legal or illegal, that are observed as illegal words, that is, $M^{-1}(M(L(G) - L))$. Reject from L the legal words for which there are illegal words observed the same way. This formula can be effectively computed with the TCT tool [14] under the hypothesis that M is a natural mask. The computation of L_R includes, however, the construction of six intermediate automata.

Cho and Marcus [5] give two algorithms for computing L_R . The first one is based on a graphical characterization of the notion of normal language. Indeed, they prove, under the assumption that R is a strict-subautomaton of G , that a language L is normal if and only if T_n is a subautomaton of T , where T and T_n are the deterministic automata for $M(L(G))$ and $M(L)$, respectively. The main step of this algorithm consists of the elimination, from T_n , of states and edges so that the resulting automaton \tilde{T}_n is the largest subautomaton of T . Language L_R equals $L \cap M^{-1}(L(\tilde{T}_n))$. In comparison, our algorithm constructs a single intermediate nondeterministic finite automaton, that is, the one for T_n . Needless to say, elimination of states and edges is performed according to a different procedure.

The second algorithm provided by Cho and Marcus [5] constructs a nondeterministic automaton for the language L_R directly from G , R , and T , under the stronger assumption that G has a property called M -recognizable. If this assumption is satisfied, the cardinality of the state set of T is less than or equal to the cardinality of the state set of G . Unfortunately, G does not always exhibit this structural property, and the computational effort required for obtaining an M -recognizable nondeterministic automaton from G is equivalent to that of their first algorithm.

III. THE ALGORITHM FOR COMPUTING $M(L_R)$

In this section, an algorithm is presented for going from R , an automaton for the legal language L , to C_2 , an automaton for the

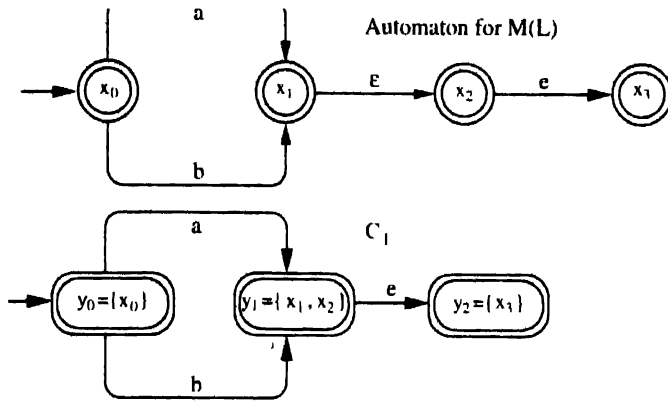


Fig. 2. Automaton for language $M(L)$, and automaton C_1 .

projection $M(L_R)$, with respect to a plant G and an observation function M . The algorithm comprises two steps.

In the first step, a deterministic automaton C_1 is constructed by applying the observation function M to every transition label of R and translating the result into a deterministic automaton. The states of C_1 are subsets of the states of R , and $L(C_1) = M(L)$. The algorithm of Lewis and Papadimitriou [8] can be used for this purpose.

Fig. 2 shows the automaton for language $M(L)$ and automaton C_1 , for the automaton R introduced in Section 1 and illustrated in Fig. 1.

Before describing the second step which computes $M(L_R)$ from G , R and M , let us state some easy-to-satisfy properties that G and R must have.

Let $T_1 = (X_1, \Sigma, f_1, x_{01}, X_1)$ and $T_2 = (X_2, \Sigma, f_2, x_{02}, X_2)$ be two deterministic automata of regular languages L_1 and L_2 , respectively, with $L_1 \subseteq L_2$. We say that T_1 refines T_2 if

$$\begin{aligned} (\forall s, t \in L_1) [f_1(s, x_{01}) = f_1(t, x_{01}) \\ \Rightarrow f_2(s, x_{02}) = f_2(t, x_{02})]. \end{aligned}$$

If T_1 refines T_2 , then it can be shown [6] that there exists a unique correspondence function $h: X_1 \rightarrow X_2$ such that

$$(\forall s \in L_1) [h \circ f_1(s, x_{01}) = f_2(s, x_{02})].$$

The correspondence function h can be intuitively interpreted as follows. Given a state x of T_1 reached after the occurrence of string s , $h(x)$ yields the state in which T_2 would be after encountering the same string.

For example, in Fig. 1, R refines G , and the correspondence function from the states of R to the states of G is

$$h(x_i) = q_i (i = 0, 1, 2, 3).$$

In the sequel,¹ let $G = (Q, \Sigma, \delta, q_0, Q)$, $R = (X, \Sigma, \xi, x_0, X)$, and $M: \Sigma \rightarrow \Lambda \cup \{\cdot\}$. We assume, without loss of generality, that R refines G . Moreover, let $h: X \rightarrow Q$ denote the correspondence function, and $C_1 = (Y_1, \Lambda, \zeta_1, y_0, Y_1)$ with $Y_1 \subseteq 2^X$.

A. The Concept of Normal Transition

A transition of C_1 , from a state y to a state y' , labeled with an event $\lambda \in (\Lambda \cup \{\cdot\})$, is normal if there is no state $x_i \in y$ and event $\tau \in \Sigma$ such that:

- in state $h(x_i)$ of the plant, event τ is active, that is, $\delta(\tau, h(x_i))$ ²;
- event τ is observed as λ , that is, $M(\tau) = \lambda$; and
- event τ is not admitted from x_i , that is, $\xi(\tau, x_i)$ is undefined.

¹In this paper, every state is marked and every language is closed.

² $\delta(\tau, h(x_i))$ means that $\delta(\tau, h(x_i))$ is defined.

Intuitively, C_1 is an intermediate structure that mirrors the legal language as observed through the mask function. A transition of C_1 labeled λ , from state y to state y' , is on the path leading to the acceptance of a word of the form $u\lambda v$, where $u, v \in \Lambda^*$. Such a word corresponds to a word in the legal language of the form $s\tau t$ with $s, t \in \Sigma^*$, $\tau \in \Sigma$, $M(s) = u$, $M(\tau) = \lambda$, and $M(t) = v$. If there exists another word $s'\tau't'$ in $L(G)$ but not in L , with $M(s') = u$, $M(\tau') = \lambda$, and $M(t') = v$, this word is illegal and observed the same way as the legal word $s\tau t$. A normal transition is one for which such a word $s'\tau't'$ does not exist. A normal transition labeled λ is called a normal λ -transition.

In the second step of the algorithm, generator C_2 is obtained from C_1 by pruning states and transitions. Pruned states are those for which nonnormal \cdot -transitions are active, and pruned transitions are those that do not have the normal property. The language of C_2 is $M(L_R)$.

The concept of normal transition is stated in an operational fashion as follows (let Σ_λ denote the set of events in Σ observed as λ , and $y \in Y_1 \subseteq 2^X$)

```
function normal (y, λ)
{precondition:  $\zeta_1(\lambda, y)!$ }
forall  $x_i \in y$  do
  forall  $\tau \in \Sigma_\lambda$  do
    if  $\delta(\tau, h(x_i))!$  and not  $\xi(\tau, x_i)!$  then
      return false
return true
```

B. The Computation of C_2

In this section, we present an algorithm for deriving a deterministic automaton C_2 from C_1 such that a word t in the language of C_2 : i) is the projection of some word s in L ; and ii) any word s' in $L(G)$ for which the projection is also t , is in L as well. More formally, the following two results will be proved (Lemmata 1, 2, 3, and Theorem 1)

$$\begin{aligned} 1) L(C_2) &= \{M(s) : s \in L \wedge (\forall s' \in s)(\forall s'' \in L(G)) \\ &\quad [M(s'') = M(s') \Rightarrow s'' \in L]\} \end{aligned}$$

and

$$2) L(C_2) = M(L_R).$$

The algorithm consists of two loops. The first one inspects every state $y \in Y_1$ of C_1 . By convention, the transition $\zeta_1(\cdot, y)$ is always defined. If this \cdot -transition is normal, then state y is copied into set Y_2 . Otherwise, the \cdot -transition is disabled by rejecting state y , because it is impossible to act on nonobservable transitions.

The second loop inspects every event $\lambda \in \Lambda$ that is active from a state $y \in Y_2$. If the transition on that event is defined in C_1 , is normal, and leads to a state in Y_2 , then it is also defined in C_2 .

If the initial state y_0 of C_1 is in Y_2 , then the algorithm returns an automaton $C_2 = (Y_2, \Lambda, \zeta_2, y_0, Y_2)$. Otherwise, it returns Φ , the empty automaton for the empty language, which has as its set of states the empty set. The computation of C_2 is performed by the following procedure

```
function step2 (M, G, R, h, C1)
{Selection of states with normal  $\cdot$ -transition}
Y2 ← {}
forall y ∈ Y1 do
  if normal(y, ·) then Y2 ← Y2 ∪ {y}
{Selection of normal λ-transitions, with λ ∈ Λ, from states in Y2}
if y0 ∈ Y2 then
```


TABLE II
COMPARISON OF THE ALGORITHMS

Alg	Authors	Restrictions	Number of intermediate automata for producing L_R	Number of intermediate automata for producing $M(L_R)$	Ease of understanding
1	Cieslak et al		3	4	Low
2	Brandt et al	M is a natural projection	6	7	High
3	Cho and Marcus (1)		4	3	Intermediate
4	Cho and Marcus (2)		3	4	Intermediate
5	Our algorithm		3	2	Intermediate

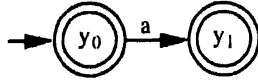


Fig. 3 Automaton C_2

```

forall  $(\lambda, y) \in (\Lambda \times Y_2)$  do
  if  $\zeta_1(\lambda, y)$  and normal( $y, \lambda$ ) and  $\zeta_1(\lambda, y) \in Y_2$  then
     $\zeta_2(\lambda, y)$  is equal to  $\zeta_1(\lambda, y)$ 
  else
     $\zeta_2(\lambda, y)$  is undefined
  return  $(Y_2 \setminus \lambda, \zeta_2, y_0, Y_2)$ 
end
return  $\Phi$ 
end
  
```

We first discuss an application of our algorithm, using the example presented in Figs. 1 and 2 then formally demonstrate its correctness.

Example Referring to Figs. 1 and 2 it is easy to see that normal(y_1) returns true for $i = 0, 1$. This is not the case for $i = 2$ however, since we have $M(f) = \epsilon$, $\delta(f, h(i_1))$ defined (that is, $\delta(f, q_1)$) and $\xi(f, i_1)$ undefined. Hence state y_2 must be eliminated as well as the incoming transition on event i .

Similarly, normal(y_0, a) is true. The result of normal(y_0, b) is false however because we have $M(c) = M(b)$, $\delta(c, h(i_0))$ defined (that is, $\delta(c, q_0)$) and $\xi(c, i_0)$ undefined. Thus the transition from y_0 to y_1 on event b must be removed. The resulting automaton C_2 is shown in Fig. 3. Another application of our algorithm can be found in Barbeau *et al.* [3].

C Demonstration of Correctness

We now formally demonstrate the correctness of our algorithm. We first introduce the following two facts.

Fact 1) By construction of C_1 given $y \in Y_1$, $i \in \Sigma$ and $s \in \Sigma^*$

$$\begin{aligned} \zeta_1(M(s), y_0) &= y \wedge i \in y \\ &\Rightarrow (\exists t \in \Sigma^*) [\xi(t, i_0) = i \wedge \delta(t, q_0) \\ &= h(i_1) \wedge M(t) = M(s)] \end{aligned}$$

Fact 2) Given $s \in I$ (that is $\xi(s, i_0)$ is defined) such that $(\forall s' \in I(G)) [M(s') = M(s) \Rightarrow s' \in I]$ if $\delta(s', q_0)$ is defined and $M(s') = M(s)$ then $\xi(s', i_0)$ is defined.

Lemma 1 Given $s \in L$,

$$\begin{aligned} (\forall s' \in L(G)) [M(s') = M(s) \Rightarrow s' \in I] \\ \Rightarrow \text{normal}(\zeta_1(M(s), y_0), \epsilon) \end{aligned}$$

Proof Since C_1 is a deterministic automaton for $M(L)$ then, for an s in L , $\zeta_1(M(s), y_0)$ is defined (it is denoted by y in the sequel).

Let $i \in \Sigma$ and $\tau \in \Sigma$ such that $\delta(\tau, h(i_1))$ is defined and $M(\tau) = M(s)$. From Fact 1) $(\exists t \in \Sigma^*) [\xi(t, i_0) = i \wedge \delta(t, q_0) = h(i_1) \wedge M(t) = M(s)]$. Thus, if we let $s' = t\tau$, $\delta(s', q_0) = \delta(t\tau, q_0) = \delta(\tau, \delta(t, q_0)) = \delta(\tau, h(i_1))$ is defined. Furthermore $M(s') = M(t\tau) = M(t)M(\tau) = M(s)$.

It follows from the hypothesis of Lemma 1 and Fact 2) that $\xi(s', i_0)$ is defined. Therefore $\xi(\tau, i) = \xi(\tau, \xi(t, i_0)) = \xi(t\tau, i_0) = \xi(s', i_0)$ is defined and by definition normal($\zeta_1(M(s), y_0)$) is true. \square

Lemma 2 Given $s \in L$

$$\begin{aligned} (\forall s' \in s)(\forall s'' \in I(G)) [M(s'') = M(s') \Rightarrow s'' \in I] \\ \Leftrightarrow M(s) \in I(C_2) \end{aligned}$$

The proof is by induction on the length $|s|$ of s .

Proof of \Rightarrow (Basis step) Let s be such that $|s| = 0$. Then $s = \epsilon$. By Lemma 1 normal($\zeta_1(M(s), y_0)$) is true. Since $M(\epsilon) = \epsilon$ and $\zeta_1(\epsilon, y_0) = y_0$ then normal(y_0) is true. By construction of C_2 $y_0 \in Y_2$. Therefore, $M(\epsilon) \in I(C_2)$.

(Induction step) Let s be such that $|s| = n + 1$ with $n \geq 0$. We may write s as $u\sigma$ for some $u \in \Sigma^*$ and some $\sigma \in \Sigma$ with $|u| = n$.

By hypothesis s is such that $(\forall s' \in s)(\forall s'' \in I(G)) [M(s'') = M(s') \Rightarrow s'' \in I]$. In particular u is such that $(\forall u' \in u)(\forall u'' \in I(G)) [M(u'') = M(u') \Rightarrow u'' \in I]$. By the induction hypothesis $M(u) \in I(C_2)$. Therefore $\zeta_2(M(u), y_0)$ is defined. For $\zeta_2(M(u\sigma), y_0)$ to be defined we have to show

- 1) that normal($\zeta_1(M(u\sigma), y_0)$) is true and
- 2) that the transition from state $\zeta_1(M(u), y_0)$ to state $\zeta_1(M(u\sigma), y_0)$ labeled $M(\sigma)$ is normal.

The first assertion follows from Lemma 1. The second assertion is verified by using an argument similar to the one developed in Lemma 1. Let $i \in \Sigma$ and $\tau \in \Sigma$ such that $\delta(\tau, h(i_1))$ is defined and $M(\tau) = M(\sigma)$. From Fact 1) $(\exists t \in \Sigma^*) [\xi(t, i_0) = i \wedge \delta(t, q_0) = h(i_1) \wedge M(t) = M(u)]$. Then if we let $s' = t\tau$, $\delta(s', q_0) = \delta(t\tau, q_0) = \delta(\tau, \delta(t, q_0)) = \delta(\tau, h(i_1))$ is defined. Furthermore, $M(s') = M(t\tau) = M(t)M(\tau) = M(u)M(\sigma) = M(u\sigma) = M(s)$. It follows from the hypothesis and Fact 2) that $\xi(s', i_0)$ is defined. Therefore $\xi(\tau, i) = \xi(\tau, \xi(t, i_0)) = \xi(t\tau, i_0) = \xi(s', i_0)$ is defined and, by definition normal($\zeta_1(M(u), y_0), M(\sigma)$) is true. By construction of C_2 $\zeta_1(M(u\sigma), y_0) \in Y_2$ and $\zeta_2(M(\sigma), \zeta_2(M(u), y_0))$ is defined. Therefore $M(s) \in I(C_2)$.

Proof of \Leftarrow (Basis step) Let s be such that $|s| = 0$. Then $s = \epsilon$ and $M(s) = \epsilon$. Let us suppose that there exists an $s'' \in I(G)$ such that $M(s'') = M(s)$. We may write $s'' = \tau_1\tau_2 \dots \tau_n$ and $M(\tau_1) = M(\tau_2) = \dots = M(\tau_n) = \epsilon$.

Since $M(s) \in I(C_2)$, then normal(y_0) is true, by construction of C_2 . Furthermore, since $\delta(\tau_1, q_0)$ is defined and $M(\tau_1) = \epsilon$, then $\xi(\tau_1, i_0)$ is defined and $\xi(\tau_1, i_0) = h^{-1}(\delta(\tau_1, q_0)) \in y_0$. In the

same way

$$\xi(\tau_2, \xi(\tau_1, x_0)) \in y_0, \dots, \xi(\tau_m, \xi(\dots, \xi(\tau_1, x_0) \dots)) \in y_0.$$

Therefore, $\xi(s'', x_0)$ is defined and $s'' \in L$.

(Induction step) Let $s \in L$, with $|s| = n + 1$ ($n \geq 0$), and $M(s) \in L(C_2)$. Since L and $L(C_2)$ are closed, for all proper prefixes s' of s , $s' \in L$, $M(s') \in L(C_2)$, and the conclusion is immediate by the induction hypothesis. It remains to be shown that the conclusion holds for $s' = s$.

We may write s as $t\sigma$ for some $t \in \Sigma^*$ and some $\sigma \in \Sigma$, with $|t| = n$. Assume that there is an $s'' \in L(G)$, with $M(s'') = M(s)$. There are two cases: $M(\sigma) = \varepsilon$ and $M(\sigma) \neq \varepsilon$.

If $M(\sigma) = \varepsilon$, then $t\sigma$ is observed as t which is a proper prefix of s and the conclusion is immediate.

Consider the case where $M(\sigma) \neq \varepsilon$. We may write $s'' = t''\tau u''$, where $t'', u'' \in \Sigma^*$, $\tau \in \Sigma$, $M(t'') = M(t)$, $M(\tau) = M(\sigma)$, and $M(u'') = \varepsilon$. Since $L(G)$ is closed, $t'' \in L(G)$ and by the induction hypothesis $t'' \in L$. Let $y = \zeta_2(M(t), y_0)$. By construction of C_2 , $\xi(t'', x_0) \in y$. Since $\zeta_2(M(t\sigma), y_0)$ is defined, then $\text{normal}(y, M(\sigma))$ is true, which implies that $\xi(t''\tau, x_0)$ is defined, and $\text{normal}(\zeta_2(M(\sigma), y), \varepsilon)$ is true, which implies that $\xi(u'', \xi(t''\tau, x_0))$ is defined. Consequently, $s'' = t''\tau u'' \in L$. \square

Lemma 3:

$$(\forall t \in \Sigma^*)[t \in L(C_2) \Rightarrow (\exists s \in L)[M(s) = t]].$$

Proof: Trivially true since C_2 is derived from C_1 , a deterministic automaton for $M(L)$, solely by pruning states and transitions. \square

Theorem 1: If L_R is the largest $(M, L(G))$ -normal sublanguage of L , then $M(L_R) = L(C_2)$.

Proof: We must show that:

- 1) if $s \in L_R$, then $M(s) \in L(C_2)$, that is, $M(L_R) \subseteq L(C_2)$; and conversely,
- 2) if $t \in L(C_2)$, then there exists an $s \in L_R$ such that $M(s) = t$, that is, $L(C_2) \subseteq M(L_R)$.

Proof of 1: Let $s \in L_R$. Since L_R is closed and $(M, L(G))$ -normal, we have that $s \in L$ and $(\forall s' \in \bar{s})(\forall s'' \in L(G))[M(s') = M(s'') \Rightarrow s'' \in L]$. From Lemma 2, we may conclude that $M(s) \in L(C_2)$.

Proof of 2: Let $t \in L(C_2)$. From Lemma 3, there exists an $s \in L$ such that $M(s) = t$, and, from Lemma 2, $(\forall s' \in \bar{s})(\forall s'' \in L(G))[M(s') = M(s'') \Rightarrow s'' \in L]$ (that is, $M^{-1}M(s') \cap L(G) \subseteq L$). From Proposition 3.3 of Cieslak *et al.* [6], we may conclude that $s \in L_R$. \square

IV. CONCLUSION

Table II gives the results of an analysis, conducted in [2], comparing the characteristics of our algorithm with those of the four other off-line algorithms mentioned in Section II.

As mentioned in the Introduction, the computation of $(M, L(G))$ -normal languages is significant because it is related to the problem of supervisory control under partial observation. The computation of L_R is an intermediate step in the solution of this problem.

Any of the algorithms listed in Table II for computing L_R can be used in combination with either of the two main algorithms for computing a controller C , namely, that of Wonham and Ramadge [15] and Cho and Marcus [5].

The algorithm in [15] is used in [6] for computing the largest $(\Delta_u, M(L(G)))$ -controllable sublanguage of $M(L_R)$ and works under the assumption that $M^{-1}(M(\Sigma_u) - \{\varepsilon\}) \subseteq \Sigma_u$. That is, no controllable event can be observed as an uncontrollable event. Furthermore, it takes as input the projection of L_R , that is, $M(L_R)$ rather than L_R . Therefore, the most suitable algorithm in that

context is the one which produces $M(L_R)$ with less effort. This is Algorithm 5 because it computes an automaton for $M(L_R)$ through the construction of only two automata.

The algorithm in Cho and Marcus [5] is more general than that in Wonham and Ramadge [15]. Indeed, the assumption that $M^{-1}(M(\Sigma_u) - \{\varepsilon\}) \subseteq \Sigma_u$ is not necessary. The former is based on the application in alternation of the algorithm of Wonham and Ramadge and an algorithm for computing an automaton that generates L_R . This process is therefore iterative, and generality is obtained at the price of a longer run-time. Any of the algorithms listed in Table II can be used for computing L_R . The most suitable one is Algorithm 4, however, because it does not require reconstruction of any intermediate automaton after the first iteration [5].

ACKNOWLEDGMENT

The authors wish to thank M. Wonham and K. Rudie for their guidance and encouragement. They would also like to thank the referees for their comments and suggestions.

REFERENCES

- [1] A. V. Aho, R. Sethi, and J. D. Ullman, *Compilers: Principles, Techniques and Tools*. Reading, MA: Addison-Wesley, 1986.
- [2] M. Barbeau, G. Couston, and R. St-Denis, "On the computation of normal languages," in *Proc. Thirty-first Annual Allerton Conf. Communication, Contr. Computing*, Univ. of Illinois at Urbana-Champaign, Sept. 1993.
- [3] —, "Requirements engineering and synthesis of a control system," *Automat. Contr. Production Syst.*, vol. 28, no. 1, pp. 37–52, 1994.
- [4] R. D. Brandt, V. Garg, R. Kumar, F. Lin, S. I. Marcus, and W. M. Wonham, "Formulas for calculating supremal controllable and normal sublanguages," *Syst. Contr. Lett.*, vol. 15, no. 2, pp. 111–117, 1990.
- [5] H. Cho and S. I. Marcus, "On supremal languages of classes of sublanguages that arise in supervisor synthesis problems with partial observation," *Mathematics Contr., Signals Syst.*, vol. 2, pp. 47–69, 1989.
- [6] R. Cieslak, C. Desclaux, A. S. Fawaz, and P. Varaiya, "Supervisory control of discrete-event processes with partial observations," *IEEE Trans. Automat. Contr.*, vol. 33, no. 3, pp. 249–260, 1988.
- [7] M. Heymann and F. Lin, "On-line control of partially observed discrete event processes with partial observation," Dept. Computer Science, Technion-Israel Institute of Technology, Haifa, Israel, Tech. Rep. CIS-9310, 1993.
- [8] H. R. Lewis and C. H. Papadimitriou, *Elements of the Theory of Computation*. Englewood Cliffs, NJ: Prentice-Hall, 1981.
- [9] F. Lin and H. Mortazavian, "A normality theorem for decentralized control of discrete event systems," *IEEE Trans. Automat. Contr.*, vol. 39, no. 5, pp. 1089–1093, 1994.
- [10] F. Lin and W. M. Wonham, "On observability of discrete-event systems," *Inform. Sciences*, vol. 44, pp. 173–198, 1988.
- [11] P. J. Ramadge and W. M. Wonham, "Supervisory control of a class of discrete event processes," *SIAM J. Contr. Optim.*, vol. 25, no. 1, pp. 206–230, 1987.
- [12] —, "The control of discrete event systems," *Proc. IEEE*, vol. 77, no. 1, pp. 81–98, 1989.
- [13] J. N. Tsitsiklis, "On the control of discrete-event dynamical systems," *Mathematics of Contr., Signals Syst.*, vol. 2, no. 1, pp. 95–107, 1989.
- [14] W. M. Wonham, "Notes on control of discrete-event systems," Dept. Electrical Engineering, University of Toronto, June 1993.
- [15] W. M. Wonham and P. J. Ramadge, "On the supremal controllable sublanguage of a given language," *SIAM J. Contr. Optim.*, vol. 25, no. 3, pp. 637–659, 1987.

Decentralized Robust Control of Uncertain Interconnected Systems with Prescribed Degree of Exponential Convergence

Zhiming Gong

Abstract—This note considers the problem of decentralized control of an uncertain interconnected time-varying system which does not satisfy the so-called matching conditions. The uncertainties, which may be decomposed into two portions, a matched portion and a mismatched portion, may appear in the subsystems and also in the interconnections between the subsystems. A robust decentralized control scheme is proposed which guarantees the controlled system to converge exponentially, with a prescribed degree, to a residue set with a prescribed bound when the mismatched portion of the uncertainties satisfies a certain bound condition.

I. INTRODUCTION

In recent years, robust control of uncertain systems has attracted great attention. Since uncertainties reside inevitably in real interconnected systems, decentralized control of uncertain interconnected systems is of great practical and theoretical interest [1]–[10].

In robust control of uncertain systems, the so-called matching conditions [11] play an important role. It is well known that an uncertain system is stabilizable when the matching conditions are satisfied (see, e.g., [11]–[14]). Many efforts have been made to relax the matching conditions on the uncertainties, and a number of sufficient conditions for existence of stabilizing controllers for uncertain systems have been obtained. Some of them are stated in terms of bounds of mismatched uncertainties [7], [15], while others are stated in terms of existence of positive definite solutions of certain Riccati equations [16].

It is interesting that the matching conditions also play an important role in decentralized control. It has been shown that when the interconnections between the subsystems of an interconnected system are through the input matrices of each subsystem, i.e., the matching condition is satisfied, the interconnected system is decentrally stabilizable [4], [5]. This remains true when the subsystems (including the input matrices) and the interconnections are assumed to have uncertainties [10] and even when the interconnections have higher-order uncertainties [9]. For the case where the matching conditions are not satisfied, Siljak [1] considered a class of interconnected systems with uncertainties in the interconnections and, by using the concept of vector Lyapunov functions and the properties of Metzler matrices, obtained a sufficient condition for stability in terms of bounds of uncertainties. Chen [8] took uncertainties in the subsystems into account and obtained a sufficient condition for stability in terms of existence of positive definite solutions of certain Riccati equations.

In this note, we consider time-varying interconnected systems with uncertainties which do not satisfy the matching conditions. The uncertainties are possibly nonlinear and fast time-varying and may appear in the subsystems (including the input matrices) and also in the interconnections among the subsystems. By decomposing the uncertainties into two portions, a matched portion and a mismatched portion, and based on the bounds of the uncertainties, we propose a robust decentralized control scheme which guarantees the controlled

system to converge exponentially, with a prescribed degree, to a residue set with a prescribed bound when the mismatched portion of the uncertainties satisfies a certain bound condition.

II. PROBLEM STATEMENT

Consider an uncertain large-scale system which is composed of N interconnected subsystems described by

$$\begin{aligned} \dot{x}_i(t) = & A_i(t)x_i(t) + B_i(t)u_i(t) + w_i(t, x_i(t), u_i(t)) \\ & + \sum_{j=1, j \neq i}^N g_{ij}(t, x_j(t)) \end{aligned} \quad (2.1)$$

where $i = 1, 2, \dots, N$; $x_i(t) \in \mathbf{R}^{n_i}$ and $u_i(t) \in \mathbf{R}^{m_i}$ are, respectively, the state and control input of the i th subsystem, and $A_i(t)$ and $B_i(t)$, which represent the nominal part of the subsystem, are matrices of appropriate dimensions and with possibly time-varying elements. The unknown functions $w_i(t, x_i(t), u_i(t))$ represent internal uncertainties in the subsystems stemming from parameter uncertainties and input disturbances, and the unknown functions $g_{ij}(t, x_j(t))$ represent uncertain interconnections between the subsystems. It is assumed that $A_i(t)$, $B_i(t)$, $u_i(t, x_i(t), u_i(t))$ and $g_{ij}(t, x_j(t))$ fulfill the conditions for existence of the solutions of (2.1).

Equation (2.1) can be rewritten routinely in the following compact form

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + w(t, x(t), u(t)) + g(t, x(t)) \quad (2.2)$$

where $x(t) = [x_1^T(t), x_2^T(t), \dots, x_N^T(t)]^T \in \mathbf{R}^n$ and $u(t) = [u_1^T(t), u_2^T(t), \dots, u_N^T(t)]^T \in \mathbf{R}^m$ are the state and control input of the overall system; $A(t)$ and $B(t)$ are block-diagonal matrices with elements $A_i(t)$ and $B_i(t)$, respectively; and $w(t, x(t), u(t))$ and $g(t, x(t))$ are the corresponding functions determined by (2.1).

Note that in system (2.1) the uncertainties represented by $w_i(t, x_i(t), u_i(t))$ and $g_{ij}(t, x_j(t))$ do not satisfy the so-called matching condition, i.e., they may not be in the range of the input matrices $B_i(t)$. Without loss of generality, it is assumed that the uncertainties can be decomposed into two portions, a matched portion and a mismatched portion, so that

$$\begin{aligned} w_i(t, x_i(t), u_i(t)) = & B_i(t)d_i^*(t, x_i(t), u_i(t)) \\ & + d_i(t, x_i(t), u_i(t)) \end{aligned} \quad (2.3a)$$

$$g_{ij}(t, x_j(t)) = B_i(t)h_{ij}^*(t, x_j(t)) + h_{ij}(t, x_j(t)) \quad (2.3b)$$

where $d_i^*(\cdot, \cdot, \cdot)$, $d_i(\cdot, \cdot, \cdot)$, $h_{ij}^*(\cdot, \cdot)$, and $h_{ij}(\cdot, \cdot)$ are unknown functions. Clearly, this decomposition is nonunique. Generally speaking, it is desirable to keep the mismatched uncertainties $d_i(\cdot, \cdot, \cdot)$ and $h_{ij}(\cdot, \cdot)$ as small as possible by an appropriate decomposition, since the matched uncertainties do not impose big difficulty in the stabilization of the system, as shown later in this note.

The following further assumptions are made on system (2.1).

Assumption 1: The pair $(A_i(t), B_i(t))$ of each subsystem is uniformly completely controllable.

Assumption 2: For each $i, j \in \{1, 2, \dots, N\}$, there exists non-negative constants ξ_{ij} , ξ_{ij}^* , ζ_i , ζ_i^* , η_{ii} , and η_{ij}^* , such that

$$\|h_{ij}(t, x_j(t))\| \leq \xi_{ij}\|x_j(t)\| + \eta_{ij} \quad (2.4a)$$

$$\|h_{ij}^*(t, x_j(t))\| \leq \xi_{ij}^*\|x_j(t)\| + \eta_{ij}^* \quad (2.4b)$$

$$\|d_i(t, x_i(t), u_i(t))\| \leq \xi_i\|x_i(t)\| + \zeta_i\|u_i(t)\| + \eta_{ii} \quad (2.4c)$$

$$\|d_i^*(t, x_i(t), u_i(t))\| \leq \xi_i^*\|x_i(t)\| + \zeta_i^*\|u_i(t)\| + \eta_{ii}^* \quad (2.4d)$$

Manuscript received November 8, 1993; revised May 12, 1994.

The author is with School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 2263.

IEEE Log Number 9408776.

with

$$\zeta_i^* < 1 \quad (2.4e)$$

where $\|\cdot\|$ denotes the standard Euclidean norm

It should be noted that Assumption 1 above specifies a condition on the nominal part of each subsystem while Assumption 2 specifies the bounds of the internal uncertainties of each subsystem and the bounds of uncertain interconnections. Condition (2.4e) can be interpreted as that the uncertainties associated with $u_i(t)$ does not override totally the nominal control effect of $u_i(t)$ to each subsystem.

To state the problem studied in this note, the following definitions are required.

Definition 1 A closed set $S \subset \mathbf{R}^n$ is called a (global uniform) final attractor of a system with the state $x(t) \in \mathbf{R}^n$ iff for each initial condition $x(t_0) = x_0$ there exists a finite constant $L(x_0, S)$ such that

$$x(t) \in S \quad \forall t > t_0 + T(x_0, S) \quad (2.5)$$

Furthermore, if the final attractor of the system is a ball $B = \{x \in \mathbf{R}^n \mid \|x\| \leq r\}$, the radius r of B is called a final bound of the system.

Definition 2 Suppose that a system with the state $x(t) \in \mathbf{R}^n$ has a final attractor B . The system is said to possess a degree of exponential convergence α toward the final attractor iff for all initial conditions $x(t_0) = x_0 \in \mathbf{R}^n \setminus B$

$$\|x(t)\| \leq f(x_0)e^{-\alpha(t-t_0)} \quad \forall x(t) \in \mathbf{R}^n \setminus B \quad \forall t > t_0 \quad (2.6)$$

where α is a positive constant and $f(x_0)$ is a positive constant possibly depending on x_0 .

The problem studied in this note can now be stated as follows: Find a decentralized state feedback control law for the system (2.1) so that the resultant closed loop system

- i) has a final attractor B with a prescribed final bound r and
- ii) possesses a prescribed degree of exponential convergence α toward the final attractor B .

III. DECENTRALIZED ROBUST CONTROL DESIGN

For system (2.1) consider first the following matrix Riccati equations

$$\begin{aligned} -\dot{P}(t) &= (A_i(t) + \alpha_i I)^T P(t) + P(t)(A_i(t) + \alpha_i I) \\ &\quad - \rho_i P(t) B_i(t) B_i^T(t) P(t) + \beta_i I \end{aligned} \quad (3.1)$$

where I denote the $n_i \times n_i$ dimensional identity matrices and α_i, ρ_i and β_i are positive numbers. Under the assumption that the pairs $(A_i(t), B_i(t))$ are uniformly completely controllable, the Riccati equations (3.1) have positive symmetrical solutions $P(t, t_f, L)$ on $[0, t_f]$ corresponding to a terminal condition $P(t_f) = L \geq 0$ [17]. Let $P_i(t) = \lim_{t_f \rightarrow \infty} P(t, t_f, 0)$. It is known that this limit exists and $P_i(t)$ are positive symmetrical solutions of (3.1) [17], [18]. Also there exist positive constants μ_i and ν_i such that [18, Lemmas 5 and 6]

$$\mu_i I \leq P_i(t) \leq \nu_i I \quad (3.2)$$

It is clear that the minimum and the maximum eigenvalues of $P_i(t)$ over all time t can serve as the positive constants μ_i and ν_i .

Now consider the following decentralized controller

$$u_i(t) = -\gamma_i B_i^T(t) P_i(t) x_i(t) \quad i = 1, 2 \quad (3.3)$$

for system (2.1), where γ_i are positive numbers. The main result of this note can then be stated in the following theorem.

Theorem 1 Consider the case where the decentralized controller (3.3) is applied to system (2.1), which satisfies the conditions given in Assumptions 1 and 2. Choose γ_i for the decentralized control (3.3) so that

$$\gamma_i > \frac{1}{2(1 - \zeta_i^* - \theta_i)} \left(\rho_i + \beta_i + \sum_{j=1}^N \xi_{ij}^* \right) \quad (3.4)$$

and

$$\zeta_i^* + \theta_i < 1 \quad (3.5)$$

where θ_i and ζ_i^* are positive numbers. Then, if there exist positive numbers δ_i and $\sigma_i, i = 1, 2, \dots, N$ such that

$$\frac{1}{2\theta_i} \zeta_i^* + \sum_{j=1}^N \frac{1}{\sigma_j} \xi_{ji}^* \leq \frac{1}{\nu_i} \left\{ r_i - \delta_i - \sum_{j=1}^N (\omega_{ij} + \xi_{ij}^*) \right\} \quad (3.6)$$

the resultant closed loop system has a final attractor B , and possesses a degree of exponential convergence α given by

$$\alpha = \min\{\alpha_1, \alpha_2, \dots, \alpha_N\} \quad (3.7)$$

toward B . The final bound r of the closed loop system is given by

$$r = \frac{1}{\delta} \left(\sigma + \sqrt{\sigma^2 + \delta^2} \right) \quad (3.8)$$

where δ, σ and r are positive numbers given by

$$\delta = \min\{\delta_1, \delta_2, \dots, \delta_N\}, \quad \sigma = \sum_{i=1}^N \left(\nu_i \sum_{j=1}^N \eta_{ij} \right)$$

$$r_i = \sum_{j=1}^N \left(\sum_{l=1}^N \eta_{lj}^* \right)$$

Proof of Theorem 1 Consider a Lyapunov function candidate

$$V(x, t) = \sum_{i=1}^N x_i^T P_i(t) x_i \quad (3.9)$$

for the overall closed loop system. It is clear from (3.2) and (3.9) that

$$\mu \|x\|^2 \leq V(x, t) \leq \nu \|x\|^2 \quad (3.10)$$

where $\mu = \min\{\mu_1, \mu_2, \dots, \mu_N\}$ and $\nu = \max\{\nu_1, \nu_2, \dots, \nu_N\}$.

The total time derivative of $V(x, t)$ along the closed loop system is given by

$$\begin{aligned} \dot{V}(x, t) &= \sum_{i=1}^N \left(x_i^T \dot{P}_i(t) x_i + x_i^T P_i(t) \dot{x}_i + \dot{x}_i^T P_i(t) x_i \right) \\ &= - \sum_{i=1}^N \left\{ 2\alpha_i x_i^T P_i(t) x_i + (2\beta_i - \rho_i) \right. \\ &\quad \times x_i^T P_i(t) B_i(t) B_i^T(t) P_i(t) x_i + \beta_i x_i^T x_i \left. \right\} \\ &\quad + \sum_{i=1}^N 2x_i^T P_i(t) \left\{ u_i(t) + \sum_{j=1, j \neq i}^N q_{ij}(t, x_j) \right\} \end{aligned} \quad (3.11)$$

By (2.3) and Assumption 2

$$\begin{aligned} &\sum_{i=1}^N 2x_i^T P_i(t) \left\{ u_i(t) + \sum_{j=1, j \neq i}^N q_{ij}(t, x_j) \right\} \\ &\leq \sum_{i=1}^N 2\|P_i(t)\| \left\{ \zeta_i \|\gamma_i B_i^T P_i(t) x_i\| + \sum_{j=1}^N (\xi_{ij} \|x_j\| + \eta_{ij}) \right\} \\ &\quad + \sum_{i=1}^N 2\|B_i^T P_i(t)\| \left\{ \zeta_i^* \|\gamma_i B_i^T P_i(t) x_i\| + \sum_{j=1}^N (\xi_{ji}^* \|x_j\| + \eta_{ji}^*) \right\} \end{aligned} \quad (3.12)$$

By using the inequality

$$2ab \leq \frac{1}{\epsilon} a^2 + \epsilon b^2 \quad (3.13)$$

for any real numbers a, b and ϵ and the fact that

$$\|P_i r_i\| \leq \nu_i \|r_i\| \quad (3.14)$$

as P_i are symmetrical matrices, it follows that

$$\begin{aligned} & \sum_{i=1}^N 2r_i^T P_i \left\{ u_i(t, r_i, u_i) + \sum_{j=1, j \neq i}^N q_{ij}(t, r_j) \right\} \\ & \leq \sum_{i=1}^N \left\{ \frac{1}{2\theta_i} \gamma_i \zeta_i^2 \nu_i^2 \|r_i\|^2 + 2\theta_i \gamma_i \|B_i^T P_i r_i\|^2 \right. \\ & \quad \left. + \sum_{j=1}^N \left(\frac{1}{\omega_{ij}} \xi_{ij}^2 \nu_j^2 \|r_j\|^2 + \omega_{ij} \|r_j\|^2 + 2\eta_{ij} \nu_j \|r_j\| \right) \right\} \\ & \quad + \sum_{i=1}^N \left\{ 2\gamma_i \zeta_i^* \|B_i^T P_i r_i\|^2 + \sum_{j=1}^N \left(\xi_{ij}^* \|B_i^T P_i r_i\|^2 \right. \right. \\ & \quad \left. \left. + \xi_{ij}^* \|r_j\|^2 + 2\eta_{ij}^* \|B_i^T P_i r_i\| \right) \right\} \quad (3.15) \end{aligned}$$

Noticing the fact that

$$\sum_{i=1}^N \sum_{j=1}^N (\omega_{ij} + \xi_{ij}^*) \|r_j\|^2 = \sum_{i=1}^N \sum_{j=1}^N (\omega_{ji} + \xi_{ji}^*) \|r_i\|^2$$

substituting (3.15) into (3.11) gives

$$\begin{aligned} V(t, r) & \leq - \sum_{i=1}^N 2\alpha_i r_i^T P_i r_i \\ & \quad - \sum_{i=1}^N \left\{ 2\gamma_i (1 - \zeta_i^* - \theta_i) - \rho_i - \sum_{j=1}^N \xi_{ij}^* \right\} \\ & \quad \times \|B_i^T P_i r_i\|^2 - \sum_{i=1}^N \left\{ \beta_i - \sum_{j=1}^N (\omega_{ji} + \xi_{ji}^*) \right. \\ & \quad \left. - \nu_i^2 \left(\frac{1}{2\theta_i} \gamma_i \zeta_i^2 + \sum_{j=1}^N \frac{1}{\omega_{ij}} \xi_{ij}^2 \right) \right\} \|r_i\|^2 \\ & \quad + \sum_{i=1}^N \sum_{j=1}^N \left(2\eta_{ij} \nu_j \|r_j\| + 2\eta_{ij}^* \|B_i^T P_i r_i\| \right) \quad (3.16) \end{aligned}$$

By the choices of γ_i and θ_i as given in (3.4) and (3.5) and by the assumption (3.6), we have

$$\begin{aligned} V(t, r) & \leq - \sum_{i=1}^N \left\{ 2\alpha_i r_i^T P_i r_i + \beta_i \|B_i^T P_i r_i\|^2 + \delta_i \|r_i\|^2 \right. \\ & \quad \left. - \sum_{j=1}^N \left(2\eta_{ij} \nu_j \|r_j\| + 2\eta_{ij}^* \|B_i^T P_i r_i\| \right) \right\} \quad (3.17) \end{aligned}$$

Since

$$\begin{aligned} & \sum_{i=1}^N \left(-\beta_i \|B_i^T P_i r_i\|^2 + \sum_{j=1}^N 2\eta_{ij}^* \|B_i^T P_i r_i\| \right) \\ & \leq \sum_{i=1}^N \frac{1}{\epsilon} \left(\sum_{j=1}^N \eta_{ij}^* \right)^2 = \epsilon \quad (3.18) \end{aligned}$$

and

$$\sum_{i=1}^N \sum_{j=1}^N \eta_{ij} \nu_j \|r_j\| \leq \sum_{j=1}^N \left(\nu_j \sum_{i=1}^N \eta_{ij} \right) \|r_j\|$$

$$\sqrt{\sum_{i=1}^N \|r_i\|^2} = \sigma \|r\| \quad (3.19)$$

it follows that

$$\begin{aligned} V(t, r) & \leq -2\alpha V(t, r) - \delta \|r\|^2 + 2\sigma \|r\| + \epsilon \\ & = -2\alpha V(t, r) - (\|r\| - r)(\delta \|r\| \\ & \quad + \sqrt{\sigma^2 + \delta\epsilon} - \sigma) \quad (3.20) \end{aligned}$$

where $\alpha, \delta, \sigma, \epsilon$ and r are positive numbers as given in (3.7) and (3.8). From (3.20), it is clear that when $\|r\| \geq r$, we have

$$V(t, r) \leq -2\alpha V(t, r) \quad (3.21)$$

and therefore

$$V(t, r) \leq V(t_0, r_0) e^{-2\alpha(t-t_0)} \quad (3.22)$$

Hence, by (3.10), we have

$$\|r(t)\| \leq \sqrt{\frac{\nu}{\mu}} \|r_0\| e^{-\alpha(t-t_0)} \quad (3.23)$$

for all $t(t) \in \mathbf{R}^n \setminus \mathbf{B}$. Equation (3.23) implies that for each initial condition $r(t_0) = r_0$, there exists a finite constant $T(t_0, \mathbf{B})$, given by

$$T(t_0, \mathbf{B}) = \frac{1}{\mu} \ln \frac{\sqrt{\nu} \|r_0\|}{\mu} \quad (3.24)$$

when $r_0 \in \mathbf{R}^n \setminus \mathbf{B}$ and $I(t_0, \mathbf{B}) = 0$ when $r_0 \in \mathbf{B}$ such that

$$r(t) \in \mathbf{B} \quad \forall t \geq t_0 + T(t_0, \mathbf{B}) \quad (3.25)$$

According to Definitions 1 and 2 (3.23) and (3.25) imply that the system has a final attractor \mathbf{B} with the final bound r and possess the degree of exponential convergence α toward the final attractor \mathbf{B} . This completes the proof of Theorem 1.

Remark 1 If the uncertainties of the subsystems and the uncertain interconnections satisfy the matching conditions, i.e. $d_i(t, r_i(t), u(t))$ and $h_{ij}(t, r_j(t))$ in (2.3) are zero and therefore ξ_{ij}, ζ_i , and η_{ij} are zero, then (3.6) can always be satisfied. In this case, the proposed decentralized control scheme can always be successful and guarantees the controlled system having a prescribed degree of exponential convergence and a prescribed final bound. For the general case, however, there is no guarantee that the condition (3.6) is satisfied, as the increasing of the numbers β_i may also cause the increasing of ν_i which are associated with the solutions $P_i(t)$ of the Riccati equations (3.1). This reflects the fact that Assumptions 1 and 2 along can not guarantee the existence of a decentralized controller which stabilizes the system. This is true even in the case of linear time-invariant systems without uncertainties [19]–[21].

Remark 2 When the free portion of the uncertainties bounded by η_{ij} and η_{ij}^* in Assumption 2, are zero, as assumed by many authors [1]–[9], a zero final bound is achieved and the system is exponentially stabilized (assuming the proposed control scheme is successful). If some numbers η_{ij} and η_{ij}^* are not zeros, however, there exists no control scheme which guarantees the controlled system to have a zero final bound, since such uncertainties can always drive the system off the nominal equilibrium at any time. In this case, large numbers γ_i and β_i may be chosen so that the final bound is satisfactorily small.

Remark 3 From (3.20), it can be seen that there exists a number $r_1 < r$ such that the derivative of $V(t, r)$ is still negative when $r_1 \leq \|r\| < r$. Therefore, the closed-loop system has a final attractor \mathbf{B}_{r_1} . The system may not possess the degree of exponential convergence α toward the final attractor \mathbf{B}_{r_1} .

Based on Theorem 1, a design procedure of the decentralized robust control scheme is given as follows:

Step 1) Verify Assumptions 1 and 2

Step 2) For a given degree of exponential convergence α , choose numbers ρ and β , and solve the matrix Riccati equations (3.1)

Step 3) According to (3.4) and (3.5) choose the numbers γ and θ and check the condition (3.6) with different sets of the positive numbers ω_i

Step 4) If in Steps 2 and 3 a set of positive numbers $\alpha, \rho, \beta, \gamma, \theta$, and ω_i is found such that (3.6) is satisfied, go to next step. Otherwise, repeat Steps 2 and 3 with different set of numbers $\alpha, \rho, \beta, \gamma, \theta$ and ω_i . If no set of the numbers $\alpha, \rho, \beta, \gamma, \theta$ and ω_i can be found such that (3.6) is satisfied, this decentralized control scheme is not applicable to the system and this system is probably not decentralized stabilizable.

Step 5) Use (3.8) to estimate the final bound γ of the closed loop system and construct the decentralized feedback controller (3.3).

It should be noted that in the computing of γ , the largest numbers α and β which satisfy (3.4) and (3.6) should be used to estimate the smallest final bound of the closed loop system. Also the design procedure given above may be repeated with the numbers $\alpha, \rho, \beta, \gamma, \theta$ and ω_i adjusted to obtain a better control law and a better estimation of the final bound.

IV. CONCLUSION

In this note, a robust decentralized control scheme for uncertain time-varying interconnected systems is proposed. General uncertainties are considered which do not satisfy the matching conditions and may appear in the subsystems and in the interconnections between the subsystems as well. The proposed control scheme is simple in the structure and yet it guarantees the controlled systems to have a prescribed degree of exponential convergence and a predictable final bound when the uncertainties satisfy a certain bound condition.

REFERENCES

- [1] D. D. Stokich, *Large Scale Systems: Stability and Structure*, Amsterdam: North Holland, 1978.
- [2] M. Jamshidi, *Large Scale Systems: Modeling and Control*, Amsterdam: North Holland, 1983.
- [3] M. K. Sundareshan, "Exponential stabilization of large scale systems: Decentralized and multilevel schemes," *IEEE Trans. Syst. Man Cybernet.*, vol. SMC-7, pp. 478-483, 1977.
- [4] M. Ikeda, O. Umefuji, and S. Kodama, "Stabilization of large scale linear systems," *Syst. Computers Contr.*, vol. 7, pp. 34-41, 1976.
- [5] M. Ikeda and D. D. Stokich, "Decentralized stabilization of linear time-varying systems," *IEEE Trans. Automat. Contr.*, vol. AC-25, no. 1, pp. 106-107, 1980.
- [6] O. Huseyin, M. F. Sizer, and D. D. Stokich, "Robust decentralized control using output feedback," *IEEE Proc.*, vol. 129, Pt. D, no. 6, pp. 310-314, 1982.
- [7] M. Ikeda and D. D. Stokich, "On optimality and robustness of LQ regulators for nonlinear and interconnected systems," in *Proc. IFAC Workshop on Model Error Concepts and Compensation*, 1985, pp. 77-82.
- [8] Y. H. Chen, "Decentralized robust control for large scale uncertain systems: a design based on the bound of uncertainty," *J. Dynamic Systems Measurement Contr.*, vol. 114, pp. 1-9, 1992.
- [9] I. Shi and S. K. Singh, "Decentralized control for interconnected uncertain systems: Extensions to higher order uncertainties," *Int. J. Contr.*, vol. 57, pp. 1453-1468, 1993.
- [10] Y. H. Chen, G. Leitmann, and X. Zhong Kai, "Robust control design for interconnected systems with time-varying uncertainties," *Int. J. Contr.*, vol. 54, no. 5, pp. 1119-1142, 1991.
- [11] S. Ghitman, "Uncertain dynamical systems—Ilyapunov min max approach," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 437-443, 1979.
- [12] M. J. Corless and G. Leitmann, "Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamic systems," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 1139-1144, 1981.
- [13] M. L. Ni and H. X. Wu, "A Riccati equation approach to the design of linear robust controllers," *Automatica*, vol. 29, pp. 1603-1605, 1993.
- [14] H. Wu and K. Mizukami, "Exponential stability of a class of nonlinear dynamical systems with uncertainties," *Syst. Contr. Lett.*, vol. 21, pp. 307-313, 1993.
- [15] B. R. Barmish and G. Leitmann, "On ultimate boundedness control of uncertain systems in the absence of matching assumptions," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 153-158, 1982.
- [16] P. P. Khargonekar, I. R. Petersen, and K. Zhou, "Robust stabilization of uncertain linear systems: quadratic stabilizability and H^∞ control theory," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 356-361, 1990.
- [17] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*, Englewood Cliffs, NJ: Prentice Hall, 1989.
- [18] M. Ikeda, H. Mieda, and S. Kodama, "Stabilization of linear systems," *SIAM J. Contr.*, vol. 10, pp. 716-729, 1972.
- [19] S. H. Wang and I. J. Davison, "On the stabilization of decentralized control systems," *IEEE Trans. Automat. Contr.*, vol. AC-18, no. 5, pp. 473-478, 1973.
- [20] M. F. Sizer and O. Huseyin, "Comments on decentralized state feedback stabilization," *IEEE Trans. Automat. Contr.*, vol. AC-26, no. 2, pp. 547-548, 1981.
- [21] Z. Gong and M. Alden, "Stabilization under decentralized information structure," in *Proc. Amer. Contr. Conf.*, 1991, pp. 922-927.

Hedging-Point Production Control with Multiple Failure Modes

Paul Glasseyman

Abstract—We consider the control of a production facility subject to multiple failure modes. Motivated by work of Akella and Kumar [1] and Bielecki and Kumar [5] on single-failure-mode models, we study hedging-point policies, in which production is controlled to its maximum rate whenever inventory is below a critical level and set to zero whenever inventory is above that level. The maximum production rate varies with the state of the machine. Assuming that the machine state is governed by a semi-Markov process, we evaluate average and discounted inventory costs for any hedging point, thus providing a simple mechanism for identifying optimal hedging points. Our most explicit results require that intervals in which demand exceeds production are exponentially distributed. We drop the exponential assumption at the expense of obtaining asymptotics rather than exact results.

I. INTRODUCTION

We consider the control of a production facility subject to various types of failures. The facility or machine can be in any of n states, labeled 1, ..., n . In state i , the machine has maximum production rate μ_i ; the production rate can be controlled to any level not exceeding the maximum rate in the current state. The durations of visits to each state are stochastic and mutually independent; those

Manuscript received September 6, 1994. This work was supported in part by NSF Grant MSS 9216490.

The author is with the Graduate School of Business, Columbia University, New York, NY 10027 USA.

IEEE Log Number 9408775.

to state i have distribution F_i . Upon the completion of a holding time in state i , the machine state becomes j with probability R_{ij} , independent of everything else. Demands arrive at a constant rate d . Production in excess of demand accumulates as inventory; when the demand rate exceeds the production rate, the inventory level decreases. Negative inventories reflect unfilled demands. Broadly, the objective is to minimize costs associated with (positive) inventory and unmet demands.

Motivated by work of Akella and Kumar [1], Bielecki and Kumar [5], and Kimemia and Gershwin [10], we analyze the performance of a simple class of control rules based on a critical number z , sometimes called a hedging point: when the inventory level is below z , produce at the maximum possible rate; when the inventory level is at z , produce at the maximum possible rate or at the demand rate, whichever is smaller; when the inventory level is above z , do not produce. Our main result (Theorem 1, below) characterizes average and discounted costs associated with this class of policies under relatively mild assumptions.

To state the result, we need to introduce some further notation. Let $r_i = d - \bar{r}_i$ be the net rate of decrease of inventory in state i , i.e., the demand rate minus the maximum production rate in that state. To rule out trivial cases, we always assume that there is at least one strictly positive r_i . We assume throughout that the matrix R of machine-state transitions is irreducible; this implies the existence of limiting probabilities (π_1, \dots, π_n) for the machine states. For each $\gamma \geq 0$, define an $n \times n$ matrix-valued function $\Phi^\gamma(\cdot)$ with entries

$$\Phi_{ij}^\gamma(\theta) = R_{ij} \int_0^\infty \exp[(r_i \theta - \gamma)x] dF_i(x). \quad (1)$$

Each $\Phi^\gamma(\theta)$ has positive entries; when these are finite, the Perron-Frobenius Theorem (see, e.g., [15]) asserts that $\Phi^\gamma(\theta)$ has a real eigenvalue $\rho_\gamma(\theta)$ equal to its spectral radius. Denote by λ_γ the strictly positive solution (whenever it exists) to

$$\rho_\gamma(\lambda_\gamma) = 1, \quad \rho_\gamma'(\lambda_\gamma) < \infty. \quad (2)$$

Suppose that cost accrues at rate $f(x)$ whenever the inventory level is x , with f mapping $(-\infty, \infty)$ to $[0, \infty)$. Denote by X_t the inventory level at time t under policy parameter z , and suppose $X_0 = z$. Let

$$V(z) = \lim_{t \rightarrow \infty} t^{-1} \int_0^t f(X_s) ds$$

be the average cost under this policy, and let

$$V_\gamma^\gamma(z) = \mathbb{E}_z \int_0^\infty e^{-\gamma t} f(X_t) dt$$

be the corresponding expected γ -discounted cost starting in machine-state i . Part of the content of our main result is that $V(z)$ is well defined.

Call state i a deficit state if $r_i > 0$. We characterize V and V_γ^γ under the assumption that holding times in deficit states are exponentially distributed. Later, we drop the exponential requirement at the expense of obtaining asymptotics and bounds, rather than exact results.

Theorem 1: Suppose that F_i is an exponential distribution for all i with $r_i > 0$.

- i) If $\sum_i \pi_i r_i < 0$ and if $\lambda_0 > 0$ solves (2) at $\gamma = 0$, then there is a $C_0 \in (0, 1)$ such that

$$V(z) = (1 - C_0)f(z) + C_0 \int_0^\infty f(z - t) \lambda_0 e^{-\lambda_0 t} dt. \quad (3)$$

- ii) For any $\gamma > 0$, if $\lambda_\gamma > 0$ solves (2), then there are constants $C_{\gamma,i}$, $i = 1, \dots, n$, such that

$$\gamma V_\gamma^\gamma(z) = (1 - C_{\gamma,i})f(z) + C_{\gamma,i} \int_0^\infty f(z - t) \lambda_\gamma e^{-\lambda_\gamma t} dt. \quad (4)$$

In proving this result, we give expressions for the constants C_0 and $C_{\gamma,i}$; these are most easily evaluated when there is a unique deficit state. We also argue that the existence of $\lambda_\gamma > 0$ in (2) should be considered typical.

A consequence of Theorem 1 is that optimal hedging points can be identified by minimizing the expressions in (3) and (4). In a case of particular interest, costs for inventory and unfilled demand are linear; i.e., there are constants $c_+, c_- \geq 0$ for which

$$f(x) = c_+ \max\{x, 0\} + c_- \max\{-x, 0\}. \quad (5)$$

Write $\log_+ x$ for $\max\{0, \log x\}$. Using (5) in (3) and (4) and minimizing over z gives the following corollary.

Corollary 1: Suppose the assumptions of Theorem 1 hold and that f has the form in (5), then

- i) under the average-cost criterion, the optimal hedging point is

$$z^* = \frac{1}{\lambda_0} \log_+ \left(\frac{C_0(c_+ + c_-)}{c_+} \right) \quad (6)$$

- ii) under the γ -discounted criterion, the optimal hedging point from initial machine state i is

$$z_{\gamma,i}^* = \frac{1}{\lambda_\gamma} \log_+ \left(\frac{C_{\gamma,i}(c_+ + c_-)}{c_+} \right). \quad (7)$$

Akella and Kumar [1] and Bielecki and Kumar [5] consider a two-state version of this model with exponentially distributed holding times and the linear cost function in (5). In that setting, they prove that the class of policies considered here includes the optimal policy under discounted and average cost criteria, respectively. They also present counterparts of (6) and (7). We do not prove (or even suggest) that these policies remain optimal in our more general setting; however, their simplicity and relative tractability makes them appealing.

The key to our analysis is a link between the inventory process and a time-reversed, continuous-time random walk in a semi-Markov environment. Through this connection, we are able to draw on results for random walks, in particular the work of Asmussen [4]. For two-state models, there is an equivalence with queues, as pointed out by Hu and Xiang [8]; see Chen and Yao [6] and Kella and Whitt [9] for related observations. None of these related references considers discounted quantities. The matrix $\Phi_0(\theta)$ and its leading eigenvalue $\rho_0(\theta)$ play an important, related role in the large deviations theory for Markov additive processes, as developed by Ney and Nummelin [14].

We prove Theorem 1 in Sections II and III. Section IV presents asymptotically optimal hedging points when holding times in deficit states have general distributions.

II. AVERAGE-COST ANALYSIS

Let us take the hedging point z to be fixed and write $X_t \equiv X_t$ for the inventory level at time t , with $X_0 = z$. Let \tilde{J}_t denote the state of the machine at time t . Our analysis of the inventory level is simplified if we work with the process $Y_t = z - X_t$, $t \geq 0$, recording the deficit in the inventory level. The process Y has the advantage that its law does not depend on z : during intervals in which $\tilde{J}_t = i$, Y_t increases at rate r_i , unless $Y_t = 0$ and $r_i \leq 0$, in which case Y_t remains at zero. This description reveals that Y is the image under reflection at the origin of a free process evolving according to the same rule but without the constraint at zero. More precisely, let $\tilde{S}_0 = 0$ and let S_t increase at rate r_i whenever $\tilde{J}_t = i$. Then

$$Y_t = \sup_{0 \leq u \leq t} (\tilde{S}_t - \tilde{S}_u). \quad (8)$$

(This type of representation is standard in queueing theory.) From this formulation, we will evaluate $V(z)$ by evaluating the expectation of $f(z - Y_t)$ with respect to the stationary distribution of Y_t .

Now let (S_t, J_t) have the law of the time-reversal of $(\tilde{S}_t, \tilde{J}_t)$, defined as follows. The process J is a semi-Markov process on $\{1, \dots, n\}$, with the same holding-time distributions F_1, \dots, F_n as \tilde{J} , and with embedded transition probabilities $R_{ij}^* = \nu_j R_{ji} / \nu_i$, where (ν_1, \dots, ν_n) are the stationary probabilities associated with the matrix R ; i.e., $\nu R = \nu$. (See, e.g., [3, Section II.5] for background on time-reversal.) In particular, then J has the same stationary distribution as \tilde{J} , given by $\pi_i = \nu_i m_i / (\sum_j \nu_j m_j)$, $i = 1, \dots, n$, with m_i the mean of F_i [7, p. 342]. The process S_t starts at zero and increases at rate r_i throughout intervals in which $J_t = i$. We take J to be right continuous.

If we give \tilde{J}_0 the stationary distribution π , then, for each $t > 0$, we obtain a stationary version of $\{J_u, 0 \leq u \leq t\}$ by setting $J_u = \tilde{J}_{t-u}$, $0 \leq u \leq t$. Similarly, we can couple S to \tilde{S} by setting $S_u = \tilde{S}_{t-u}$, $0 \leq u \leq t$. With this construction, from (8) we get

$$Y_t = \sup_{0 \leq u \leq t} (S_t - S_u) = \sup_{0 \leq u \leq t} (\tilde{S}_t - \tilde{S}_{t-u}) = \sup_{0 \leq u \leq t} S_u.$$

Therefore, for any $x > 0$ and any $i, j \in \{0, 1, \dots, n\}$, the events $\{Y_t \leq x, \tilde{J}_0 = i, \tilde{J}_t = j\}$ and $\{\sup_{0 \leq u \leq t} S_u \leq x, J_0 = j, J_t = i\}$ coincide, and so have the same probability. Defining $M_t = \sup_{0 \leq u \leq t} S_u$ and using the stationarity of J and \tilde{J} , we conclude that

$$\pi_i P_i(Y_t \leq x, J_t = j) = \pi_j P_j(M_t \leq x, J_t = i) \quad (9)$$

where the subscripts on P indicate the initial machine state (J_0 on the left, J_t on the right). By summing over i , we get

$$\sum \pi_i P_i(Y_t \leq x, J_t = j) = \pi_j P_j(M_t \leq x) \quad (10)$$

Next, we consider the limit as t increases. Because M_t is almost surely increasing in t , the limit $M \triangleq \lim_{t \rightarrow \infty} M_t$ exists with probability one; moreover, under the drift condition $\sum_i \pi_i r_i < 0$ in Theorem 1-i), M is almost surely finite [2, p. 309]. Also, since (Y_t, J_t) is evidently regenerative, it has a limiting distribution not depending on the initial state; let (Y_∞, J_∞) have that limiting distribution. Then from (10) we get

$$\begin{aligned} P(Y_\infty \leq x, J_\infty = j) &= \lim_{t \rightarrow \infty} \sum \pi_i P_i(Y_t \leq x, J_t = j) \\ &= \lim_{t \rightarrow \infty} \pi_j P_j(M_t \leq x) \\ &= \pi_j P_j(M \leq x). \end{aligned} \quad (11)$$

Summing over j , we have proved the following (compare [3, Proposition 2.2]).

Lemma 1: Under the assumptions in Theorem 1-i), $P(Y_\infty \leq x) = \sum_j \pi_j P_j(M \leq x)$. Consequently, $V(z) = E[f(z - Y_\infty)] = \sum_j \pi_j E_j[f(z - M)]$.

In light of Lemma 1, to prove Theorem 1-i) it suffices to evaluate the distributions $P_j(M \leq \cdot)$, $j = 1, \dots, n$ and to show that the expression in the lemma coincides with (3). The properties we need can in principle be obtained from the analytical results of [2]. Instead, we use a change-of-measure argument based on similar techniques in [4], [13], [17]. We work with the discrete-time process $\{(S_{\tau_n}, J_{\tau_n}), n = 0, 1, \dots\}$, where τ_n is the epoch of the n th jump of J .

For fixed $\gamma \geq 0$, let λ_γ be as in (2). We need two preliminary results. They are easily verified by induction and similar to results in [4] so we omit their proofs.

Lemma 2: The quantity $E_j[\exp(\lambda_\gamma S_{\tau_n} - \gamma \tau_n); J_{\tau_n} = i]$ is the ji -entry of $(\Phi^\gamma(\lambda_\gamma))^n$.

By the Perron-Frobenius theorem, $\Phi^\gamma(\lambda_\gamma)$ has a strictly positive right-eigenvector h (depending on γ) associated with the maximal eigenvalue $\rho_\gamma(\lambda_\gamma) \equiv 1$; i.e., $\Phi^\gamma(\lambda_\gamma)h = h$. We have the following lemma.

Lemma 3: $E_j[\exp(\lambda_\gamma S_{\tau_n} - \gamma \tau_n)h(J_{\tau_n})/h(j)] = 1$, for all j and n .

A consequence of Lemma 3 is that, with $J_0 = j$, $\exp(\lambda_\gamma S_{\tau_n} - \gamma \tau_n)h(J_{\tau_n})/h(j)$ defines a change of measure. More explicitly, with

$$\tilde{F}_i(x) = \frac{Rh(i)}{h(i)} \int_0^x e^{(\lambda_\gamma - \gamma)u} F_i(du) \quad (12)$$

and $Rh(j) = \sum_i R_{ij}h(j)$, \tilde{F}_i is a probability distribution function. Let \tilde{P} and \tilde{E} denote probability and expectation, respectively, when the holding times in state i have distribution \tilde{F}_i , $i = 1, \dots, n$. These are related to the original process through a version of Wald's identity.

Lemma 4: For any $\gamma \geq 0$, suppose there is a $\lambda_\gamma > 0$ solving $\rho_\gamma(\lambda_\gamma) = 1$. If N is a stopping time for the process $\{(S_{\tau_n}, J_{\tau_n}), n \geq 0\}$ and if the event A is measurable with respect to $\{(S_{\tau_n}, J_{\tau_n}), 0 \leq n \leq N\}$, then $P_i(A; N < \infty) = E_i[\exp\{-(\lambda_\gamma S_{\tau_N} - \gamma \tau_N)\}h(i)/h(J_{\tau_N}); N < \infty]$.

Similar results are proved in [4], [13], and [17]. The proof is standard, so we omit it.

Under the new measure, S has positive drift.

Lemma 5: When the holding times have distributions F_i , $i = 1, \dots, n$, $\lim_{t \rightarrow \infty} t^{-1} S_t > 0$ (and is independent of the initial state).

Proof: By Lemma 5.3 of [14], $\lim_{t \rightarrow \infty} t^{-1} S_t = \rho_\gamma(\lambda_\gamma)$. From [15, Theorem 3.7], we know that $\rho_\gamma(\cdot)$ is convex. We now separate the cases $\gamma = 0$ and $\gamma > 0$. For the former, we find that $\Phi^0(0) = R$, a stochastic matrix, so $\rho_0(0) = 1$. Moreover, $\rho'_0(0)$ is the drift of S under the original measure and is therefore negative, by hypothesis. Thus, the convexity of ρ_0 implies that if $\lambda_0 > 0$ exists, then ρ_0 is increasing at λ_0 . In the case $\gamma > 0$, we see that $\Phi^\gamma(0)$ is strictly substochastic and therefore $\rho_\gamma(0) < 1$. Again, convexity implies that ρ_γ must be increasing at λ_γ . \square

Remark. The convexity of the functions $\rho_\gamma(\cdot)$, $\gamma \geq 0$, indicates that the existence of λ_γ solving (2) is typical, at least if the deficit-state holding-time distributions have exponential tails. Indeed, the only alternative then to $\rho_\gamma(\lambda_\gamma) = 1$ is the existence of a $\theta > 0$ for which $\rho_\gamma(\theta) < 1$ and $\rho_\gamma(\theta) = \infty$. Such cases must be considered exceptional; see [2] for a treatment of this case. The second part of (2) makes the new drift finite.

We can now evaluate the distribution of M from any initial state. Let $N_x = \inf\{n \geq 0 : S_{\tau_n} > x\}$ be the index of the first jump epoch at which S exceeds x . (In a simplifying abuse of notation, we write S_{N_x} and J_{N_x} for $S_{\tau_{N_x}}$ and $J_{\tau_{N_x}}$, and we write J_{N_x-} for the state of the machine when S first exceeds x .) Under the original (negative-drift) measure, N_x may be infinite, but under the new (positive-drift) measure, N_x is almost surely finite for all $x \geq 0$. By Lemma 4

$$\begin{aligned} P_j(M > x) &\equiv P_j(N_x < \infty) \\ &= \tilde{E}_j[\exp(-\lambda_0 S_{N_x})h(j)/h(J_{N_x})] \\ &= e^{-\lambda_0 x} h(j) \tilde{E}_j[\exp(-\lambda_0 [S_{N_x} - x])/h(J_{N_x})]. \end{aligned}$$

To evaluate the expectation on the right, we condition on J_{N_x-} , which is necessarily a deficit state. Given $J_{N_x-} = k$, the overshoot $S_{N_x} - x$ and the next state J_{N_x} are independent. Moreover, F_k exponential implies F_k exponential, so the overshoot is exponentially distributed with mean $r_k m_k$, where

$$m_k = \frac{Rh(k)}{h(k)} \int_0^\infty x e^{-\lambda_0 x} F_k(dx) = \frac{m_k}{1 - \lambda_0 r_k m_k} \quad (13)$$

is the mean of \tilde{F}_k . With $J_0 = k$, S_{τ_1} also has the exponential distribution \tilde{F}_k . Thus, we have shown that

$$\begin{aligned} P_j(M > x) &= h(j) \sum \tilde{P}_j(J_{N_x-} = k) \tilde{E}_k[\exp(-\lambda_0 S_{\tau_1})] \tilde{E}_k[1/h(J_{\tau_1})] e^{-\lambda_0 x} \\ &= h(j) \sum_k \tilde{P}_j(J_{N_x-} = k) \frac{1}{1 - \lambda_0 r_k m_k} Rh^{-1}(k) e^{-\lambda_0 x}. \end{aligned} \quad (14)$$

From (13) we find that $1/(1 + \lambda_0 r_k \bar{m}_k) = 1 - \lambda_0 r_k m_k$. So, setting

$$C_0 = \sum_j \pi_j h(j) \sum_k \bar{P}_j(J_{N_x-} = k) (1 - \lambda_0 r_k m_k) R h^{-1}(k)$$

we find from (11) and (14) that $P(Y_\infty > x) = C_0 e^{-\lambda_0 x}$, for all $x \geq 0$. Since $V(z) = E[f(z - Y_\infty)]$, this completes the proof of part (i) of Theorem 1.

The only unknown terms in our expression for C_0 are the probabilities $\bar{P}_j(J_{N_x-} = k)$. If there is just one deficit state—call it state 1—then

$$C_0 = \sum_j \pi_j h(j) (1 - \lambda_0 r_1 m_1) R h^{-1}(1). \quad (15)$$

Each of the terms appearing in this expression is easily computed through matrix calculations if the number of machine states is not too large—not more than 30, say. Choosing h so that $\pi h = 1$ eliminates the first factor on the right in (15).

III. DISCOUNTED-COST ANALYSIS

We now turn to the evaluation of $V_i^\gamma(z)$, $\gamma > 0$. As in the previous section, we take z to be fixed and suppress it as an argument. Let L be an exponentially distributed random variable with mean $1/\gamma$, independent of everything else. Then

$$\gamma V_i^\gamma = \gamma E_i \int_0^\infty e^{-\gamma t} f(X_t) dt = E_i[f(X_L)].$$

Thus, to evaluate V_i^γ , it suffices to find the distribution of the inventory level at the random time L . As in the previous section, we work with the process $Y_t = z - X_t$ rather than the inventory level itself, and now seek to evaluate

$$\gamma V_i^\gamma = E_i[f(z - Y_L)]. \quad (16)$$

From (9) we get

$$\begin{aligned} P_i(Y_L > x) &= \frac{1}{\pi_i} \sum_j \pi_j P_j(M_L > x, J_L = i) \\ &= \frac{1}{\pi_i} \sum_j \pi_j P_j(T_i < L, J_L = i) \end{aligned} \quad (17)$$

where $T_i = \inf\{t \geq 0 : S_t \geq x\}$ is the first time S reaches level x . Letting $L' = L - T_i$, we get

$$\begin{aligned} P_j(T_i < L, J_L = i) &= P_j(T_i < L) P_j(J_L = i | T_i < L) \\ &= E_j[e^{-\gamma T_i}] P_j(J_{T_i+L'} = i | T_i < L). \end{aligned}$$

Given $\{T_i < L\}$, L' is exponentially distributed with mean $1/\gamma$. Thus, we get

$$\begin{aligned} P_j(T_i < L, J_L = i) &= E_j[e^{-\gamma T_i}] \sum_k P_j(J_{T_i} = k | T_i < L) \\ &\quad \times P_k(J_L = i). \end{aligned}$$

We conclude from (17) that

$$\begin{aligned} P_i(Y_L > x) &= \frac{1}{\pi_i} \sum_j \pi_j E_j[e^{-\gamma T_i}] \\ &\quad \times \sum_k P_j(J_{T_i} = k | T_i < L) P_k(J_L = i). \end{aligned} \quad (18)$$

We evaluate $E_j[e^{-\gamma T_i}]$ by adapting a change-of-measure argument due to Glynn (personal communication) and Kollman [11]. Let $\gamma > 0$ be fixed and let h denote a strictly positive Perron-Frobenius right-eigenvector for $\Phi^\gamma(\lambda_\gamma)$, so that $\Phi^\gamma(\lambda_\gamma)h = h$. Via Lemma 4, define new holding-time distributions as specified by (12). Throughout this section \bar{P} and \bar{E} refer to probability and expectation based on these distributions. Let $N_t = \inf\{n \geq 0 : S_{\tau_n} > x\}$, as before.

Lemma 6: $E_j[e^{-\gamma T_i}] = b_j e^{-\lambda_\gamma x}$, where

$$b_j = h(j) \sum_k \bar{P}_j(J_{N_x-} = k) (1 + (\gamma - r_k \lambda_\gamma) m_k) R h^{-1}(k).$$

Proof: Because $T_i \leq \tau_{N_x}$, a.s., and because the evolution of S_t is deterministic between jumps, T_i is measurable with respect to $\{(S_{\tau_n}, J_{\tau_n}), 0 \leq n \leq \tau_{N_x}\}$; so, we may evaluate the expectation of $\exp(-\gamma T_i)$ by applying a measure transformation to $\{(S_{\tau_n}, J_{\tau_n}), 0 \leq n \leq N_x\}$ as follows

$$\begin{aligned} E_j[e^{-\gamma T_i}] &= \bar{E}_j[e^{-\gamma T_i} e^{\gamma \tau_{N_x} - \lambda_\gamma x_{N_x}} h(J_{N_x}) / h(J_{N_x-})] \\ &= e^{-\lambda_\gamma x} h(j) \bar{E}_j[e^{\gamma(\tau_{N_x} - T_i)} e^{-\lambda_\gamma(x_{N_x} - x)} / h(J_{N_x-})] \\ &= e^{-\lambda_\gamma x} h(j) \bar{E}_j[e^{(\gamma - \lambda_\gamma \tau_{N_x-})(\tau_{N_x} - T_i)} / h(J_{N_x-})] \end{aligned} \quad (19)$$

where we have written r_{N_x-} for $r_{J_{N_x-}}$ (the prevailing net rate when S first crosses x) and used the fact that $(S_{N_x} - x) = r_{N_x-}(\tau_{N_x} - T_i)$, a.s. Now we condition on J_{N_x-} , which is necessarily a deficit state. The corresponding holding-time distribution is therefore exponential; and given J_{N_x-} , the next state J_{N_x} is independent of $\tau_{N_x} - T_i$. Thus

$$\begin{aligned} E_j[e^{-\gamma T_i}] &= e^{-\lambda_\gamma x} h(j) \sum_k \bar{P}_j(J_{N_x-} = k) \bar{P}_k[e^{(\gamma - r_k \lambda_\gamma)\tau_1}] \\ &\quad \cdot \bar{E}_k[1/h(J_{\tau_1})] \\ &= e^{-\lambda_\gamma x} h(j) \sum_k \bar{P}_j(J_{N_x-} = k) \\ &\quad \cdot R h^{-1}(k) \\ &\quad \cdot [1 - (\gamma - r_k \lambda_\gamma) \bar{m}_k] \\ &= e^{-\lambda_\gamma x} h(j) \sum_k \bar{P}_j(J_{N_x-} = k) \\ &\quad \times (1 + (\gamma - r_k \lambda_\gamma) m_k) R h^{-1}(k), \\ &= b_j e^{-\lambda_\gamma x} \end{aligned}$$

where we have used

$$\begin{aligned} \bar{m}_k &= \frac{R h(k)}{h(k)} \int_0^\infty x e^{-(\gamma + r_k \lambda_\gamma)x} F_k(dx) \\ &= \frac{1}{1 + (\gamma - \lambda_\gamma r_k) m_k} \end{aligned} \quad \square$$

If we set

$$C_{\gamma,i} = \frac{1}{\pi_i} \sum_j \pi_j b_j \sum_k P_j(J_{T_i} = k | T_i < L) P_k(J_L = i) \quad (20)$$

then from (18) and Lemma 6, we find that $P_i(Y_L > x) = C_{\gamma,i} e^{-\lambda_\gamma x}$. Using this in (16) concludes the proof of part (ii) of Theorem 1.

As in the average-cost setting, a more explicit expression for $C_{\gamma,i}$ is available if we assume that only state 1 is a deficit state. With this assumption, $J_{N_x-} = 1$, a.s., and the sums over k in (18) and (20) collapse, resulting in

$$\begin{aligned} C_{\gamma,i} &= \frac{1}{\pi_i} \left(\sum_j \pi_j h(j) \right) (1 + (\gamma - r_1 \lambda_\gamma) m_1) \\ &\quad \times R h^{-1}(1) P_1(J_L = i). \end{aligned} \quad (21)$$

To conclude the evaluation of this constant, we need to find $g_i(j) \triangleq P_j(J_L = i)$ at $j = 1$. For $\gamma > 0$, let D_γ be the diagonal matrix with entries

$$D_\gamma(i,i) = 1 - \sum_j \Phi_{ij}^\gamma(0).$$

Let e_i be the n -dimensional vector $(0, \dots, 0, 1, 0, \dots, 0)'$ with i th component equal to one.

Lemma 7 $q_i = (I - \Phi(0))^{-1} D_i c_i$

Proof Notice that

$$q_i(t) \equiv P_j(J_t = i) = \int_0^t \mathbf{1}_{\{J_t = i\}} dt$$

The result now follows by combining Proposition 2.20 and (5.14) of Chapter 10 of Çinlar [7], noting that the Q_γ defined in his (2.10) coincides with our $\Phi^\gamma(0)$ \square

Remarks

- i) As $\gamma \downarrow 0$, $P_j(J_t = i) \rightarrow \tau_i$ for all i, j and we expect that $\lambda_\gamma \rightarrow \lambda_0$. Thus $C_\gamma \rightarrow C_0$ for all i , and (3) and (4) are consistent as γ decreases to zero.
- ii) Under the discounted cost criterion, the optimal hedging point depends on the initial state, denote by γ_i the optimal hedging point starting in machine state i . An alternative policy controls the system to hedging point γ_i whenever the machine state is i . Such a policy is guaranteed to result in lower cost than following a single γ . The optimal policy of Akella and Kumar [1] in a two state model is precisely of this form, however since in their model the hedging point for the breakdown state is always greater than that for the functional state, the former plays no role, and the policy is indistinguishable from one with a single hedging point. See Kimemia and Geishwin [10], Malhamé [12] and Sharifnia [16] for more on state dependent hedging points.

IV. ASYMPTOTICS FOR GENERAL DEFICIT INTERVALS

We now drop the requirement that the holding time distributions for deficit states be exponential at the expense of obtaining asymptotics rather than exact results. For the asymptotics we restrict attention to the linear cost structure in (5) and consider optimal hedging points as $c \rightarrow \infty$. The case of large penalties for unfilled demands is of practical as well as theoretical interest.

For the following recall that a distribution is nonlattice if it is not concentrated on any set of the form $\{x - 2\delta - \delta \leq x \leq 2\delta - \delta\}$.

Theorem 2 Suppose that F is nonlattice whenever $i > 0$.

- i) If $\sum \tau_i < 0$ and if $\lambda_0 > 0$ solves (2) at $c = 0$ then there is a constant $C_0 \in (0, 1)$ such that as $c \rightarrow \infty$

$$C_\gamma \sim \left(\frac{C_0(c_+ + c_-)}{c_+} \right) + o(1) \quad (22)$$

- ii) Suppose there is a unique deficit state. For any $\gamma > 0$, $\lambda_\gamma > 0$ solves (2) then there are constants C_γ , $\gamma = 1$ such that as $c \rightarrow \infty$

$$C_\gamma \sim \frac{(c_+ + c_-)}{c_+} + o(1) \quad (23)$$

Proof The argument used for Theorem 1 still applies, except that $S_{\lambda_\gamma} - I$ and $\tau_{\lambda_\gamma} - I$ no longer have exponential distributions. In the average cost setting the argument leading to (15) now shows that

$$P(\lambda_\gamma > i) = \sum \pi_j h(j) E_j [e^{-\lambda_\gamma (S_{\lambda_\gamma} - i)} / h(\Gamma_{\lambda_\gamma})] \quad (24)$$

Set $B_i = S_{\lambda_\gamma} - i$ and $\Gamma_i = \Gamma_{\lambda_\gamma}$ and consider the process $\{(B_i, \Gamma_i) : i \geq 0\}$, with i playing the role of time parameter. This process is regenerative, the regeneration points are those i for which $S_{\lambda_\gamma} = i$ and $\Gamma_{\lambda_\gamma} = j$ for some fixed j . It follows (under our nonlattice condition) that this process has a limiting distribution not depending on the initial state, let $(B_\infty, \Gamma_\infty)$ have that distribution. Then, by the very definition of convergence in distribution, the limit

$$\lim E_j [e^{-\lambda_0 (S_{\lambda_\gamma} - i)} / h(\Gamma_{\lambda_\gamma})] = E[e^{-\lambda_0 B_\infty} / h(\Gamma_\infty)]$$

exists. We have thus shown that with

$$C_0 = \sum \pi_j h(j) E[e^{-\lambda_0 B_\infty} / h(\Gamma_\infty)]$$

we have

$$P(\lambda_\gamma > i) \sim C_0 e^{-\lambda_0 i} \quad (25)$$

the symbol \sim indicating that the ratio of the two expressions converges to one as $i \rightarrow \infty$ (see also [2]). So long as the distribution of λ_γ is continuous on $(0, \infty)$, straightforward minimization of $V(-) = c_+ E[(S - \lambda_\gamma)^+] + c_- E[(\lambda_\gamma - S)^+]$ shows that the optimal hedging point satisfies

$$C_\gamma = 1 \quad (26)$$

The required continuity follows from the nonlattice assumption. But then (25) implies that

$$\frac{1}{C_\gamma} e^{-\lambda_0 i} \left(\frac{c_+}{c_+ + c_-} \right) \rightarrow 1$$

which implies (22). The proof of (23) is similar but starts from the representation

$$P(\lambda_\gamma > i) = \frac{1}{\pi} \sum_j \tau_j h(j) E_j [e^{-(\lambda_\gamma - S_{\lambda_\gamma})(\tau_{\lambda_\gamma} - i)} / h(\Gamma_{\lambda_\gamma})] \times P_j(J_t = i | T < I) e^{-\lambda_\gamma i}$$

When state 1 is the only deficit state, this reduces to

$$P(\lambda_\gamma > i) = \frac{1}{\pi} \sum_j \tau_j h(j) R h^{-1}(1) E_j [e^{-(\lambda_\gamma - S_{\lambda_\gamma})}] \times P_1(J_t = i) e^{-\lambda_\gamma i} \quad (27)$$

much as in Section III. We claim that $-\lambda_\gamma i_1 \leq 0$. Otherwise, all row sums of $\Phi(\lambda_\gamma)$ would be strictly less than one and since the maximum row sum is an upper bound on the spectral radius we would arrive at the contradiction $\rho(\lambda_\gamma) < 1$. By essentially the same regenerative argument used above for C_0 we may define

$$C_\gamma = \frac{1}{\pi} \sum_j \tau_j h(j) R h^{-1}(1) \times P_1(J_t = i) \lim_{i \rightarrow \infty} E_j [e^{-(\lambda_\gamma - S_{\lambda_\gamma})(\tau_{\lambda_\gamma} - i)}]$$

and conclude that $P_i(\lambda_\gamma > i) \sim C_\gamma e^{-\lambda_\gamma i}$. The rest of the proof is the same as part i) \square

It is possible that part ii) of this result extends to multiple deficit states. Such an extension would require convergence of $P_j(J_t = i | T < I)$ as $i \rightarrow \infty$ and a guarantee that $\tau - \lambda_\gamma i_1 \leq 0$ whenever $i_1 > 0$.

V. CONCLUDING REMARKS

- i) For simplicity, we have assumed throughout that the time spent in a machine state is independent of the next state visited. One can easily imagine settings in which it would be desirable to relax this requirement. For example, a quick repair might restore the machine to operation sooner than a thorough repair, but the resulting machine states might differ in their susceptibility to further failure. To incorporate such phenomena, one could define holding-time distributions F_{ij} corresponding to current-state i and next-state j . The matrix $\Phi(\theta)$ is defined just as in (1) but with F_i replaced by F_{ij} . The analysis goes through as before, except that in expressions like (14)

$$\frac{1}{1 + \lambda_0 r_k m_k} \sum R_k h^{-1}(i)$$

becomes

$$\sum R_{k,i} h^{-1}(i) / (1 + \lambda_0 r_k \bar{m}_{k,i})$$

with $\bar{m}_{k,i}$ the mean of $F_{k,i}$.

- ii) It also seems possible to carry out an extension in which the hedging point changes with the machine state, with one important modification. Let Z_t be the hedging point at time t , a function of \tilde{J}_t . It is possible that upon a change in the machine state the inventory level X_t exceeds the new hedging point Z_t . Under the usual operation of the system, X_t would then decrease linearly at the rate of demand until it reaches Z_t . Suppose we modify the system so that X_t is instantaneously reduced to Z_t whenever a change in machine state results in $X_t > Z_t$; physically, this corresponds to discarding excess inventory rather than waiting to sell it. With this modification, the process $Y_t = Z_t - X_t$ becomes the image under the reflection mapping of the free process $Z_t - \tilde{S}_t$, and its distribution can be analyzed through time reversal.

ACKNOWLEDGMENT

The author would like to thank P. Glynn for introducing him to measure transformations for discounted costs.

REFERENCES

- [1] R. Akella and P. R. Kumar, "Optimal control of production rate in a failure prone manufacturing system," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 116-126, 1986.
- [2] K. Arndt, "Asymptotic properties of the distribution of the supremum of a random walk on a markov chain," *Th. Probab. Appl.*, vol. 25, pp. 309-324, 1980.
- [3] S. Asmussen, *Applied Probability and Queues*. Chichester: Wiley, 1987.
- [4] —, "Risk theory in a markovian environment," *Scand. Actuarial J.*, pp. 69-100, 1989.
- [5] T. Bielecki and P. R. Kumar, "Optimality of zero-inventory policies for unreliable manufacturing systems," *Oper. Res.*, vol. 36, pp. 532-541, 1988.
- [6] H. Chen and D. D. Yao, "A fluid model for systems with random disruptions," *Oper. Res.*, vol. 40, pp. S239-S247, 1992.
- [7] E. Çinlar, *Introduction to Stochastic Processes*. Englewood Cliffs, NJ: Prentice-Hall, 1975.
- [8] J. Q. Hu and D. Xiang, "The queueing equivalence to a manufacturing system with failures," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 499-502, 1993.
- [9] O. Kella and W. Whitt, "A storage model with a two-state random environment," *Oper. Res.*, vol. 40, pp. S257-S262, 1992.
- [10] J. Kimemia and S. B. Gershwin, "An algorithm for the computer control of a flexible manufacturing system," *IEE Trans.*, vol. 15, pp. 353-362, 1983.
- [11] C. Kollman, "Deep-penetration calculations for scattering neutrons by importance sampling," Statistics Dept., Stanford Univ., Stanford, CA, working paper, 1993.
- [12] R. P. Malhamé, "Ergodicity of hedging control policies in single-part multiple-state manufacturing systems," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 34-343, 1993.
- [13] H. D. Miller, "A generalization of wald's identity with applications to random walks," *Ann. Math. Statist.*, vol. 32, pp. 549-560, 1961.
- [14] P. Ney and E. Nummelin, "Markov additive processes. I. eigenvalues and limit theorems. II. large deviations," *Ann. Probab.*, 15, pp. 561-609, 1986.
- [15] E. Seneta, *Non-Negative Matrices*. New York: Wiley, 1973.
- [16] A. Sharifnia, "Production control of a manufacturing system with multiple machine states," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 620-625, 1988.
- [17] M. C. K. Tweedie, "Generalizations of wald's fundamental identity in sequential analysis to markov chains," *Proc. Cambridge Philos. Soc.*, 56, pp. 205-214, 1960.

A Globally Optimal Minimax Solution for Spectral Overbounding and Factorization

Robert E. Scheid and David S. Bayard

Abstract—In this paper, an algorithm is introduced to find a minimum phase transfer function of specified order whose magnitude "tightly" overbounds a specified real-valued nonparametric function of frequency. This method has direct application to transforming nonparametric uncertainty bounds (available from system identification experiments and/or plant modeling) into parametric representations required for modern robust control design software (i.e., a minimum-phase transfer function multiplied by a norm-bounded perturbation).

I. INTRODUCTION

Assume that a discrete-time plant $P(z^{-1})$ is estimated as $\hat{P}(z^{-1})$, and let L denote the uncertainty in the estimate. For example, three common characterizations of plant uncertainty are L_A -additive uncertainty, L_I -input multiplicative uncertainty, and L_O -output multiplicative uncertainty, where [17, p. 224]

$$\begin{aligned} L_A &= P - \hat{P} \\ L_I &= \hat{P}^{-1}(P - \hat{P}) \\ L_O &= (P - \hat{P})\hat{P}^{-1} \end{aligned} \quad (1)$$

Note that multiplicative representations require a square plant. Let L denote any one of the above three quantities. Suppose, a nonparametric overbound $\ell(\omega)$ on L is known such that

$$\ell(\omega) > \bar{\sigma}(L(e^{-j\omega T})) \quad \text{for all } \omega \in [0, \pi/T] \quad (2)$$

where T is the sampling period and $\bar{\sigma}(L)$ is the maximum singular value of L . Various methods are available to find $\ell(\omega)$ from raw data (cf., [3], [11], [13], [16]). $\ell(\omega)$, however, is a nonparametric function of frequency and cannot be used directly in modern robust control software packages such as the Matlab Robust Control Toolbox [7] and μ synthesis software [2]. Instead, the uncertainty must be represented as a minimum phase transfer function matrix $\mathcal{W}(z^{-1})$ of a specified order such that

$$L(\epsilon) = \Delta \mathcal{W}(\epsilon) \quad (3)$$

where Δ is norm-bounded, i.e.,

$$\|\Delta\|_\infty < 1.$$

The choice of \mathcal{W} in (3) can be structured or unstructured. For present purposes, the simplest choice is to use a scalar matrix representation

$$\mathcal{W} = W \cdot I \quad (4)$$

where W is a single-input single-output rational function.

Manuscript received December 21, 1993. This work was supported in part by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

The authors are with the Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109 USA.

IEEE Log Number 9508782.

To incorporate the uncertainty bound (3) into a robust control design, a systematic method for determining the weighting W in (4) is needed. Ideally, any approach to find W should satisfy the following properties:

P1) W must overbound the uncertainty ℓ , i.e.

$$\|W(e^{-j\omega T})\| \geq \ell(\omega) \quad \text{for all } \omega \in [0, \pi/T]$$

to ensure the existence of some $\|\Delta\|_\infty < 1$ satisfying (3)

P2) W should be as tight an overbound as possible to avoid conservatism in the final robust control design

P3) W should be of specified order (in fact as low order as possible) since it will be incorporated as a weighting and increase the final controller order

P4) W should be stable and minimum phase

These requirements on W rule out using several powerful methods from the complex analysis literature. For example, interpolation methods such as Nevanlinna-Pick theory for finding rational interpolants to complex valued data [20] do not address the real valued data case and do not satisfy the above properties. For the same reasons, the extension of interpolatory methods to the noisy data case (cf. [13], [14], [12], [6], [18], etc.) are not directly applicable to the present problem. Even if interpolatory methods could be appropriately modified to overbound real valued data sets (i.e., P1)) and do so in a minimax optimal sense (i.e., P2)) the final rational fits tend to be of very high order (e.g., on the order of the number of data points) and would not satisfy properties P3) and P4) in general.

In this paper, a mathematical programming approach is used to find a W which satisfies all of the properties P1)–P4). The main idea is to reformulate the problem so as to find a spectrally factorizable rational function W^*W whose magnitude tightly overbounds the squared data $\ell^2(\omega)$. This is done in Section II by posing a minimax nonlinear optimization problem to ensure tightness of fit of a rational function having specified order, with side constraints to ensure that the data is overbounded and that the solution admits a spectral factorization. A key result in Section III is that the nonlinear optimization problem can be solved by a sequence of reweighted constrained linear problems. In particular, a globally convergent linear programming spectral overbounding and factorization (LPSOF) algorithm is presented based on solving a sequence of linear programming problems [19]. The LPSOF algorithm provides a globally optimal solution to the nonlinear problem.

II. PROBLEM FORMULATION

In this section, a nonlinear constrained optimization is posed to compute a minimum phase transfer function W of order m such that $\|W\|$ is a tight overbound on $\ell(\omega)$ for all ω . With this result, the uncertainty can be written in standard form $I = \Delta W$ where $\|\Delta\|_\infty < 1$. Forming the quantity $W(e^{-j\omega T})W^*(e^{-j\omega T})$ and evaluating on the unit circle gives an expression of the form

$$W^*W = f(\omega) \quad (5)$$

where

$$f(\omega) = f_0 + f_1 \cos(\omega T) + \dots + f_m \cos(m\omega T) \quad (6a)$$

$$\alpha(\omega) = 1 + \alpha_1 \cos(\omega T) + \dots + \alpha_m \cos(m\omega T) \quad (6b)$$

It is noted that $\alpha(\omega)$ is defined as monic without loss of generality.

A. Constraints for Overbounding

The requirement that $\|W\|$ be an overbound on $\ell(\omega)$ is equivalent to the requirement that $\|W\|^2$ is an overbound on ℓ^2 and can be

expressed as

$$\geq \ell^2(\omega) \quad \text{for all } \omega \in [0, \pi/T] \quad (7)$$

B. Constraints for Tight Overbounding

The requirement that $\|W\|^2$ be a 'tight' overbound can be expressed as

$$\min_{\delta} \delta \quad (8)$$

where

$$\delta \geq \left\{ \left(\frac{\ell^2(\omega)}{\alpha(\omega)} - \ell^2(\omega) \right) q^{-1}(\omega) \right\} \quad \text{for all } \omega \in [0, \pi/T] \quad (9)$$

Here, the criterion minimizes a worst case error δ which is frequency weighted by the quantity $q^{-1}(\omega)$.

C. Constraints for Spectral Factorizability

The requirement that the overbound f/α admits a spectral factorization can be satisfied by ensuring that (Astrom [1])

$$f(\omega)/\alpha(\omega) > 0 \quad \text{for all } \omega \in [0, \pi/T] \quad (10a)$$

$$\alpha(\omega) > 0 \quad \text{for all } \omega \in [0, \pi/T] \quad (10b)$$

Note that condition (10a) is implied by (7) and condition (10b) can be enforced explicitly by the constraint

$$\alpha(\omega) > 0 \quad \text{for all } \omega \in [0, \pi/T] \quad (11a)$$

for some small α . For technical reasons, it will be convenient to enforce a similar constraint on f as

$$f(\omega) > f_{\min} > 0 \quad \text{for all } \omega \in [0, \pi/T] \quad (11b)$$

for some small f_{\min} .

In summary, it is desired to solve the optimization problem (8), (9) for α, f subject to constraints (7) and (10a)–(10b).

III. THE LPSOF ALGORITHM

In this section, the LP spectral overbounding and factorization (LPSOF) algorithm is introduced which solves the constrained nonlinear optimization problem of Section II on grid of points $\Lambda = \{\omega_1, \dots, \omega_n\}$. Modifications to extend these results to all $\omega \in [0, \pi/T]$ will also be discussed.

The constrained optimization problem restricted to points of the set Λ can be written as

$$\min_{\delta} \delta \quad (12)$$

subject to

$$f(\omega_i) - \ell^2(\omega_i)\alpha(\omega_i) \geq 0 \quad (13a)$$

$$f(\omega_i) - \ell^2(\omega_i)\alpha(\omega_i) \leq \delta q(\omega_i)\alpha(\omega_i) \quad (13b)$$

$$f(\omega_i) \geq f_{\min}, \alpha(\omega_i) \geq \alpha_{\min} \quad (13c)$$

for all $\omega_i, i = 1, \dots, n$

where $\alpha(\omega)$ and $f(\omega)$ are defined by (6a), (6b). A key observation from (12), (13) is that for fixed δ , the optimization over α, f is simply a linear programming problem to find a feasible solution for the coefficients α, f . Hence, the joint optimization problem can be solved by a nested search procedure where an outer loop systematically decreases δ , while an inner loop finds feasible solutions in the variables α and f for fixed δ . The procedure terminates when the smallest δ is found, which admits a feasible solution. This

approach is denoted as the LP spectral overbounding and factorization (LPSOF) algorithm.

To solve problems (12) and (13), one must begin with upper and lower bounds for the optimal value δ . For example, one can choose the lower bound $\delta_- = 0$ and let the upper bound δ_+ be derived from some starting feasible suboptimal solution (an obvious choice is $\alpha = 1$, $\beta = \max_{\omega_i} t^2(\omega_i)$). Then $\hat{\delta} = (\delta_+ + \delta_-)/2$ becomes an updated value for δ_+ or δ_- depending on whether or not the inequalities (13) can be satisfied for $\delta = \hat{\delta}$ (i.e., the bisection method [8]). In this way, the LPSOF algorithm converges to the optimal value of δ geometrically (i.e., as a power of 1/2).

This process can be further accelerated by effectively linearizing about the candidate value $\hat{\delta}$. Thus, given $\hat{\delta}$, one can solve the linearized problem

$$\max_{\alpha, \beta, \delta} u \quad (14)$$

subject to

$$\beta(\omega_i) - t^2(\omega_i)\alpha(\omega_i) \geq 0 \quad (15a)$$

$$\beta(\omega_i) - t^2(\omega_i)\alpha(\omega_i) - \hat{\delta}q(\omega_i)\alpha(\omega_i) \leq -u \quad (15b)$$

$$\beta(\omega_i) \geq \underline{\beta}, \quad \alpha(\omega_i) \geq \underline{\alpha} \quad (15c)$$

for all $\omega_i, i = 1, \dots, n$.

Then $\hat{\delta}$ provides an update for δ_- or δ_+ according to whether the solution u is negative or nonnegative. In the latter case a sharper *a-posteriori* estimate for δ_+ is derived via

$$\delta_+ = \max_{\omega \in \Lambda} \left\{ \left(\frac{\beta(\omega)}{\alpha(\omega)} - t^2(\omega) \right) q^{-1}(\omega) \right\} \quad (16)$$

where $\alpha(\omega)$ and $\beta(\omega)$ are the solutions derived from (14), (15). It is also worth noting that setting $\hat{\delta} = 0$ and solving (14)–(16) provides an excellent starting value for δ_+ to initialize the algorithm.

Remark 1. It is noted that the weighting q may be chosen as certain functions of the unknown polynomials α and β without violating the linear form of the constraints. Generally $q(\omega)$ may be taken in the form

$$q(\omega) = q_0(\omega) + q_1(\omega) \frac{\beta(\omega)}{\alpha(\omega)} \quad (17)$$

where $q_0(\omega)$ and $q_1(\omega)$ are positive for $\omega \in \Lambda$ and specified before hand.

Remark 2: By the fundamental properties of linear programming [9], [10], the LPSOF algorithm is globally convergent and achieves a globally optimal solution to the discrete problem (12), (13). An excellent algorithm for spectrally factorizing polynomials β and α without solving for roots can be found in Kucera [21].

Remark 3: For each fixed value of δ the semi-infinite linear programming problem (7), (9), and (11) (i.e., for continuous-valued $\omega \in [0, \pi/T]$) can be solved as a sequence of discretized linear programming problems of the form (13) in the limit as the mesh becomes sufficiently fine (cf. [15, Section 7.2]). Thus, one can recover the solution to the semi-infinite spectral overbounding and factorization problem (7), (8), (9), and (11) using the LPSOF approach, by solving a sequence of linear programs with decreasing mesh spacing for each fixed value of δ . Decreasing the mesh spacing, however, may not be desirable in practice, and alternative methods for approximating the solution to the semi-infinite problem will be presented in Section IV.

IV. MODIFICATIONS OF THE LPSOF ALGORITHM

Strictly speaking, the LPSOF algorithm only enforces inequalities (7), (10a), (10b) at the points in the grid Λ . Hence, the inequalities may be violated in between grid points, and the solution may not

be a true overbound, and/or may not admit a spectral factor. If this happens in practice, the simplest solution, generally, is to choose a denser grid (see Remark 3) and/or increase lower bounds $\underline{\alpha}$, $\underline{\beta}$ in (13c). There may be certain cases, however, where these approaches are not desirable. Hence, in this section, systematic modifications of the LPSOF algorithm are presented to overcome this problem.

For convenience to subsequent discussion, we make the following assumptions.

Assumption 1: Let $t^2(\omega)$ be a linear spline interpolant to the points $t^2(\omega_i)$ defined on the grid $\Lambda = \{\omega_1, \dots, \omega_n\}$ with piecewise linear segments having maximum slope κ , and maximum grid size $h = \max_i \{\omega_{i+1} - \omega_i\}$.

Assumption 2: The set of vectors $\{\cos(k\omega_1 T), \dots, \cos(k\omega_n T)\} \in \mathbb{R}^n$, $k = 0, \dots, m$ are linearly independent.

Modifications of the LPSOF algorithm to ensure proper behavior between grid points, fall into two categories, *a-priori* and *a-posteriori*. These methods will be discussed separately below.

A. A-Priori Modifications

The basic idea behind the *a-priori* modifications is to enforce additional linear constraints in the LPSOF algorithm so that the derivatives $\alpha'(\omega) = \frac{d}{d\omega}(\alpha(\omega))$, $\beta'(\omega) = \frac{d}{d\omega}(\beta(\omega))$ and $(\beta/\alpha)' = \frac{d}{d\omega}(\beta/\alpha)$ are suitably bounded for all $\omega \in [0, \pi/T]$. This clearly restricts the excursions of α , β , and β/α in between grid points, so that under Assumption 1, and specification of lower bounds $\underline{\alpha}$, $\underline{\beta}$ in (13c), the desired inequalities (7), (10a), (10b) can be satisfied.

Some useful definitions are in order: A function $x(\omega)$ defined on the interval $\omega \in \Omega$ is said to be uniformly bounded from above if $|x(\omega)| \leq C' < \infty$ for all $\omega \in \Omega$. Here, the quantity C' is denoted as the uniform upper bound. Similarly, the function $x(\omega)$ is said to be uniformly bounded from below if $|x(\omega)| \geq c' > 0$ for all $\omega \in \Omega$. The quantity c' is denoted as the uniform lower bound.

A method to uniformly bound the aforementioned derivatives is now introduced. For some positive $\bar{\alpha}$, $\bar{\Lambda}$, K_α , and K_β , let the following linear constraints be imposed on the grid Λ

$$\alpha(\omega_i) \leq \bar{\alpha} \quad (18a)$$

$$\beta(\omega_i) \leq K_\alpha \alpha(\omega_i) \quad (18b)$$

$$|\alpha'(\omega_i)| \leq K_\alpha |\alpha(\omega_i)| \quad (18c)$$

$$|\beta'(\omega_i)| \leq K_\beta |\beta(\omega_i)| \quad (18d)$$

If linear constraints (18a)–(18d) are used to augment linear constraints (13a)–(13c) of the LPSOF algorithm, it can be shown that

$$\underline{\alpha} \leq \alpha(\omega_i) \leq \bar{\alpha} \quad (19a)$$

$$\underline{\beta} \leq \beta(\omega_i) \leq K_\alpha \bar{\alpha} \quad (19b)$$

$$|\alpha'(\omega_i)| \leq K_\alpha \bar{\alpha} \quad (19c)$$

$$|\beta'(\omega_i)| \leq K_\beta K_\alpha \bar{\alpha} \quad (19d)$$

$$d \left(\frac{\beta}{\alpha} \right) \leq K' (K_\beta + K_\alpha) \quad (19e)$$

where (19e) follows from

$$\left(\frac{\beta}{\alpha} \right)' = \frac{\beta'}{\alpha} - \frac{\alpha' \beta}{\alpha^2} = \frac{\beta'}{\alpha} \left(\frac{\beta'}{\beta} - \frac{\alpha'}{\alpha} \right). \quad (20)$$

Inequalities (19a)–(19e) imply that α , β , β/α and their derivatives are bounded on the grid Λ . Under Assumption 2, these bounds at the grid points impose bounds on the coefficients α_i , β_i (this is because the matrix of trigonometric functions which determines these coefficients has a bounded inverse, (cf. [8])). This implies that $|\alpha(\omega)|$, $|\beta(\omega)|$, $|\alpha'(\omega)|$ and $|\beta'(\omega)|$ are uniformly bounded from above since they are bounded functions of the bounded coefficients

α_i, β_i . If the uniform upper bounds on $|\alpha'(\omega)|$ and $|\beta'(\omega)|$ are sufficiently small, constraints (13c) on the grid points imply the existence of uniform (nonzero) lower bounds on $|\alpha(\omega)|, |\beta(\omega)|$. Hence, by systematically decreasing the values of K_i and K_{ii} (with the other constraints fixed), one can make $|\alpha'(\omega)|$ and $|\beta'(\omega)|$ arbitrarily small, and there will always be some point at which inequalities (10a), (10b) are satisfied uniformly in ω . Given that (10a), (10b) is satisfied in this manner, it follows from (20) that $|(j/\alpha)'|$ is bounded uniformly from above. Using (13a) and Assumption 1, it follows that (7) is satisfied when this uniform upper bound on $|(j/\alpha)'|$ falls below the value of κ defined in Assumption 1.

In summary, by augmenting the linear constraints (13a)–(13c) of the LPSOF algorithm by linear constraints (18a)–(18d), and solving a sequence of problems where K_i and K_{ii} are systematically decreased, there will be a point at which (7) (10a), (10b) are satisfied, ensuring that β/α is an overbound on t^2 for all $\omega \in [0, \pi/T]$ and spectrally factorizable.

B. A-Posteriori Modifications

The basic idea behind *a-posteriori* modifications is to slightly perturb the unmodified LPSOF solution so that inequalities (7), (10a), (10b) are satisfied uniformly in ω . For grid Λ sufficiently fine, a small perturbation can always be found which does the job. To see this, note that the solution to problems (12), (13) will always satisfy bounds of the form (19), if the quantities α, K, K_{ii} , and K_i are computed *a-posteriori*. As noted in the previous discussion, Assumption 2 guarantees corresponding bounds for the coefficients α_i, β_i which in turn impose uniform upper bounds on $|\alpha(\omega)|, |\beta(\omega)|, |\alpha'(\omega)|$ and $|\beta'(\omega)|$. Then, for a sufficiently small grid size h , uniform constraints (7), (10a), (10b) can be satisfied by means of an $O(h)$ perturbation of $\alpha(\omega), \beta(\omega)$, and δ .

The construction above implicitly assumes that α, K, K_{ii} , and K_i are reasonably sized. If not, these quantities can be explicitly constrained *a-priori*, as done earlier.

V. NUMERICAL EXAMPLE

In this section, the LPSOF algorithm is used to determine spectrally factorizable overbounds on additive uncertainty estimates obtained from large space structure identification experiments [5]. Raw additive uncertainty data $t^2(\omega_i)$ adapted from [5] is shown in Fig. 1, depicted by the symbol ** , on the grid $\Lambda = \{\omega_i = i\pi/(128T), i = 1, \dots, 128\}$, where $T = .05$ seconds. In addition to the raw data set, the envelope \bar{t}^2 is depicted in Fig. 1. The envelope \bar{t}^2 is a smoothed nonparametric overbound on the raw data t^2 .

In all subsequent examples, the LPSOF algorithm of Section III is used in sequential linearized form (14), (15), with underbounds $\alpha_i = \beta_i = 0$ and frequency weighting $q(\omega) = q_0 + q_1 \beta(\omega)/\alpha(\omega)$, where $q_1 = 1$ and q_0 remains to be specified (see (17)). The dual rather than primal form of the underlying LP problems [9], [10] is implemented to considerably reduce the number of constraint equations. In all runs, 10 to 15 iterations were sufficient to ensure convergence of δ to six significant digits.

A. Overbounding Raw Data

In this section, the raw data t^2 of Fig. 1 is overbounded using the LPSOF algorithm of Section III.

The first set of runs is generated by fixing $m = 4$ and $q_1 = 1$ and varying weighting factor q_0 as $q_0 = 0, .01, .1, 1$. The results are summarized in Fig. 2, where the raw data t^2 is depicted by the symbol ** . It is seen that the overbound associated with $q_0 = 0$ does a reasonably good job of overbounding the data, but tends to sacrifice some accuracy in the peaks for accuracy in the troughs (e.g., there

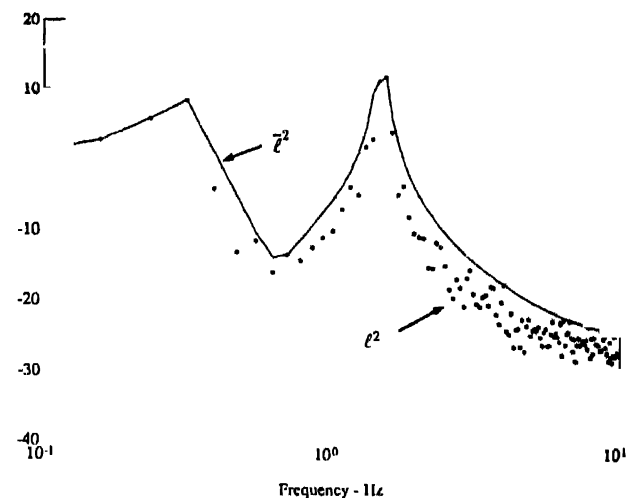


Fig. 1 Raw additive uncertainty data (t^2 (denoted by **) and envelope data (\bar{t}^2 (denoted by *)) defined on 128 point grid.

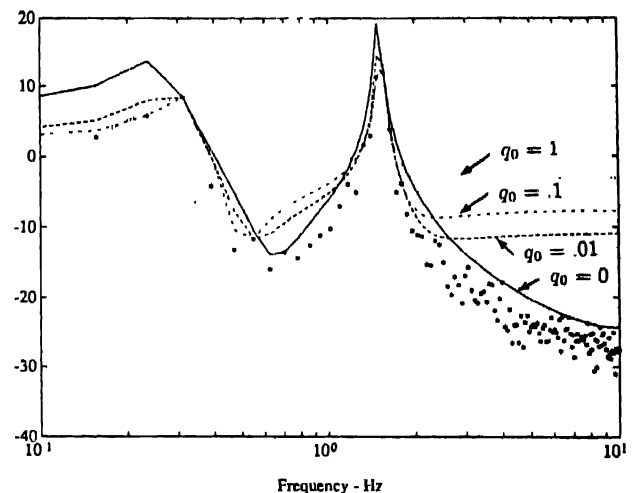


Fig. 2 Spectrally factorizable overbounds β/α on raw data t^2 (denoted by **) using unmodified LPSOF algorithm ($\alpha_i = \beta_i = 0, q_1 = 1, m = 4$) obtained by varying weighting factor $q_0 = 0, .01, .1, 1$

is about an 8 db overshoot of the main peak). Methods to improve the results become apparent.

Method 1: Deemphasize the weighting of trough data by increasing q_0 .

Method 2: Reduce oscillatory behavior of the data by overbounding the envelope \bar{t}^2 rather than raw data t^2 .

Method 1 motivates the remaining runs of Fig. 2 which are generated by increasing q_0 . (Method 2 motivates using envelope data rather than raw data, which will be discussed later.) It is seen from Fig. 2, that increasing q_0 from 0 to .01 improves the accuracy of the overbound by several db in the vicinity of the peaks at the cost of accuracy in the troughs. Increasing q_0 further to .1 and 1 continues this trend. Hence, the user can choose from the family of curves in Fig. 2, to trade-off accuracy in the peaks for accuracy in the troughs.

The second set of runs is generated by fixing $q_1 = 1, q_0 = 1$ and varying the order m as $m = 2, 4, 6$. The results are summarized in Fig. 3. It is seen that the successive overbounds improve uniformly as the order is increased. In particular, the peaks are fitted reasonably well by bounds of all orders (this is a consequence of using $q_0 = 1$ for all runs) while most of the improvement from increasing order comes from fitting the troughs.

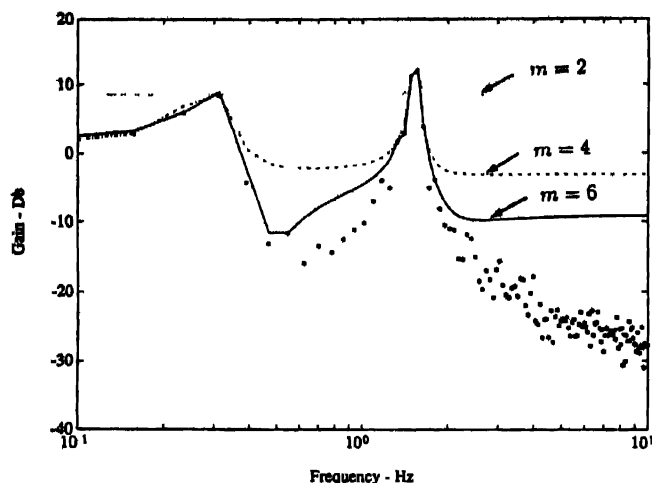


Fig 3 Spectrally factorizable overbounds $1/\alpha$ on raw data l^2 (denoted by '*') using unmodified LPSOF algorithm ($\alpha = \beta = 0$, $q_0 = 1$, $q_1 = 1$) obtained by varying bound order $m = 2, 4, 6$

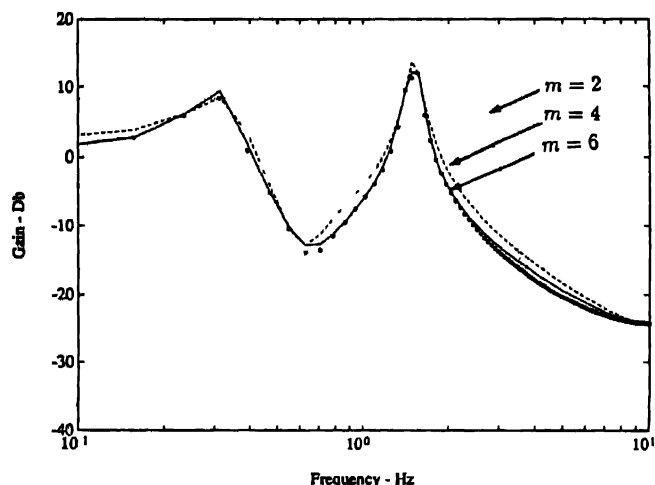


Fig 4 Spectrally factorizable overbounds $1/\alpha$ on envelope data \tilde{l}^2 (denoted by '*') using unmodified LPSOF algorithm ($\alpha = \beta = 0$, $q_0 = 0$, $q_1 = 1$) obtained by varying bound order $m = 2, 4, 6$

B Overbounding Envelope Data

In this study (in accordance with Method 2 outlined above), oscillatory data is avoided by overbounding the envelope data \tilde{l}^2 rather than raw data l^2 (see Fig 1).

The runs of Fig 4 are generated by fixing $q_0 = 0$, $q_1 = 1$ and varying the order m as $m = 2, 4, 6$. The envelope data \tilde{l}^2 is depicted by the symbol '*'. It is seen that the overbounds for $m = 4$ and $m = 6$ are excellent, and in fact the latter nearly interpolates the data.

VI. CONCLUSION

A systematic method, denoted as the LPSOF algorithm, has been developed for finding a minimum-phase transfer function of specified order whose magnitude "tightly" overbounds a specified nonparametric real-valued function of frequency. The main idea is to find a spectrally factorizable rational function which tightly overbounds the data "squared." This leads to a nonlinear constrained optimization problem which can be solved by a sequence of linear programming problems.

The original motivation behind the development of the LPSOF algorithm was to systematically replace the graphical overbounding method used in [4] for determining robust control weightings. The algorithm is, however, generally useful for determining spectral factors from raw PSD data and can be useful in such applications as deconvolution, disturbance identification, blind channel equalization, and estimation of noise coloring filters.

REFERENCES

- [1] K. J. Astrom, *Introduction to Stochastic Control Theory*, New York: Academic, 1970.
- [2] G. J. Balas, J. C. Doyle, K. Glover, A. K. Packard, and R. Smith, *H-Infinity and Mu Control Analysis: Mu-Tools Manual* (beta test version), Sept. 1990.
- [3] D. S. Bayard, "Statistical plant set estimation using Schroeder-phased multisinusoidal input design," *1 Applied Mathematics and Computation*, vol. 58, pp. 169-198, 1993.
- [4] D. S. Bayard, Y. Yam, and E. Mettler, "A criterion for joint optimization of identification and robust control," *IEEE Trans Automat Contr*, vol. 37, no. 7, pp. 986-991, July 1992.
- [5] D. S. Bayard, F. Y. Hadeagh, Y. Yam, R. E. Scheid, E. Mettler, and M. H. Milman, "Automated on-orbit frequency domain identification for large space structures," *Automatica*, vol. 27, no. 6, pp. 931-946, Nov. 1991.
- [6] J. Chen, C. N. Nett, and M. K. H. Fan, "Worst-case system identification in H_∞ : Validation of *a priori* information, essentially optimal algorithms, and error bounds," in *Proc Amer Contr Conf*, Chicago, June 1992, pp. 251-257.
- [7] R. Y. Chiang and M. G. Safanov, *Robust Control Toolbox*, The Math Works Inc., 1988.
- [8] G. Dahlquist and A. Bjork, *Numerical Methods*, Englewood Cliffs, NJ: Prentice-Hall, 1974.
- [9] G. B. Dantzig, *Linear Programming and Extensions*, Princeton, NJ: Princeton Univ. Press, 1963.
- [10] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, New York: Academic, 1981.
- [11] G. C. Goodwin and M. E. Salgado, "Quantification of uncertainty in estimation using an embedding principle," in *Proc Amer Contr Conf*, Pittsburgh, PA, June 21-23, 1989.
- [12] G. Gu and P. P. Khargonekar, "A class of algorithms for identification in H_∞ ," *Automatica*, vol. 28, no. 2, pp. 299-312, 1992.
- [13] A. Helmicki, C. A. Jacobson, and C. N. Nett, " H_∞ Identification of stable LSI systems: A scheme with direct application to controller design," in *Proc Amer Contr Conf*, Pittsburgh, PA, 1989.
- [14] A. J. Helmicki, C. A. Jacobson, and C. N. Nett, "Control oriented system identification: A worst-case/deterministic approach in H_∞ ," *IEEE Trans Automat Contr*, vol. 36, pp. 1163-1176, Oct. 1991.
- [15] R. Hetlich and K. O. Kortanek, "Semi-infinite programming: Theory, methods and applications," *SIAM Review*, vol. 15, no. 3, pp. 380-429, Sept. 1973.
- [16] R. L. Kosut, "On-line identification and control tuning of large space structures," in *Proc Fifth Yale Conf Adaptive Systems Theory*, Yale University, May 1987.
- [17] M. Morari and E. Zafriou, *Robust Process Control*, Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [18] J. R. Partington, "Robust identification and interpolation in H_∞ ," *Int J Contr*, vol. 54, no. 5, pp. 1281-1290, 1991.
- [19] R. E. Scheid, D. S. Bayard, and Y. Yam, "A linear programming approach to characterizing norm-bounded uncertainty from experimental data," in *Proc Amer Contr Conf*, Boston, MA, June 1991, pp. 1956-1958.
- [20] N. Young, *An Introduction to Hilbert Space*, Cambridge: Cambridge Univ. Press, 1988.
- [21] V. Kucera, *Discrete Linear Control: The Polynomial Approach*, New York: Wiley, 1979.

A Lower Bound for Limiting Time Delay for Closed-Loop Stability of an Arbitrary SISO Plant

R. Devanathan

Abstract—This correspondence is concerned with the time delay margin for closed-loop stability of a single input, single output plant with time delay uncertainty. Since no restriction is placed on the stabilizing controller, the limiting value of time delay depends on the plant transfer function only. The approach, using a modified form of classical Nevanlinna–Pick interpolation theory, provides a lower bound for the limiting time delay given the plant open-loop poles and zeros only.

NOTATION

\mathbb{C} = {Complex plane}
 $A = \{s \in \mathbb{C}, \text{Re}(s) > 0\}$
 $\bar{A} = \{s \in \mathbb{C}, \text{Re}(s) \geq 0\}$
 $\hat{A} = \bar{A} \cup \{\infty\}$
 $U = \{s \in \mathbb{C}, |s| < 1\}$
 $\bar{U} = \{s \in \mathbb{C}, |s| \leq 1\}$

I. INTRODUCTION

The closed-loop stability of single-input single-output (SISO) systems with time delay uncertainty is considered in this paper. Walton and Marshall [1] have given a direct method for stability analysis of time delay systems. Consider the closed-loop system of Fig. 1. Let the process transfer function be given by

$$p(s) = p_0(s) \exp(-T_d s), \quad T_d \geq 0 \quad (1.1)$$

where $p_0(s)$ corresponds to the nominal transfer function and T_d corresponds to the time delay uncertainty. Walton and Marshall [1] consider the following problem. Given $p_0(s)$ and the controller transfer function $c(s)$, what is the limiting (minimum) value of time delay T_d below which the closed-loop system of Fig. 1 is stable? El-Sakkary [2] has given a partial solution to this problem.

Suppose we now ask: Consider the closed-loop system of Fig. 1 where the plant $p(s)$ can be any transfer function of the form of (1.1). What is the largest value of the limiting (minimum) time delay, call it $T_{d\max}$, below which there exists a controller stabilising the closed-loop system of Fig. 1? This is known as the optimum time delay margin problem. In this paper, we attempt to characterize a lower bound for $T_{d\max}$.

II. BACKGROUND

Let the nominal real rational transfer function $p_0(s)$ correspond to m finite zeros, $q_i, i = 1, 2, \dots, m$, multiplicities included, and n poles, $p_j, j = 1, 2, \dots, n$, multiplicities included, in the right half of the complex plane including the imaginary axis, i.e., in \bar{A} . The problem is well posed [3] when $p_0(s)$ is strictly proper.

Let the zero at infinity of $p_0(s)$ be denoted by q_∞ , of multiplicity $(n - m)$. Let $E(s)$ be such that

$$E(s) = p_0(s) c(s) / \{1 + p_0(s) c(s)\}. \quad (2.1)$$

Given $E(s)$, (2.1) can be used to find $c(s)$. The closed-loop system of Fig. 1 corresponding to the nominal system is stable if and only if [4]

- i) $E(s)$ is analytic in \bar{A}

Manuscript received June 2, 1993; revised June 14, 1994.

The author is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 2263.

IEEE Log Number 9408783.

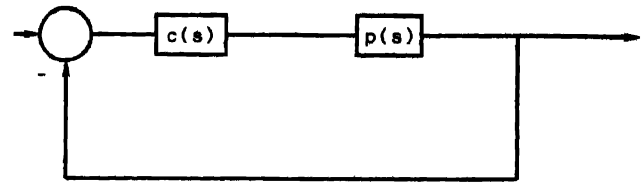


Fig. 1. Closed-loop system.

- ii) $E(q_i) = 0, i = 1, 2, \dots, m$, multiplicities included
- iii) $E(q_\infty) = 0$, multiplicities included
- iv) $E(p_j) = 1, j = 1, 2, \dots, n$, multiplicities included.

(2.2)

Consider the conformal mapping [5] as shown in Fig. 2. \bar{U} is the closed unit disk while U corresponds to the open unit disk. G can be any simply connected region in $\mathbb{C} \cup \{\infty\}$ containing the points 0 and 1.

ϕ is a mapping such that

$$\phi: G \rightarrow U \quad (2.3)$$

and

$$\phi(0) = 0. \quad (2.4)$$

The conditions ii), iii), and iv) of (2.2) correspond to

- i) $E(\infty) = 0, i = 1, 2, \dots, m$, multiplicities included
- ii) $E(z_{q_\infty}) = 0$, multiplicities included
- iii) $E(\infty) = \phi(1), j = 1, 2, \dots, n$, multiplicities included

(2.5)

where

$$E: \bar{U} \rightarrow U \quad (2.6)$$

$$\psi: \bar{A} \rightarrow \bar{U} \quad (2.7)$$

$$z_{q_i} = \psi(q_i), \quad i = 1, 2, \dots, m \quad (2.8)$$

$$z_{q_\infty} = \psi(q_\infty) \quad (2.9)$$

$$z_{p_j} = \psi(p_j), \quad j = 1, 2, \dots, n. \quad (2.10)$$

Following [5], consider the mapping E of (2.6) such that

$$\bar{E}(z_t) = \alpha b_t, \quad t = 1, 2, \dots, 2n, \quad \alpha \geq 0 \quad (2.11)$$

where

$$z_t = \infty, t = 1, 2, \dots, m, \text{ and } t = 1, 2, \dots, m, \text{ respectively,} \quad (2.12)$$

$$z_t = z_{q_\infty}, t = m + 1, \dots, n \quad (2.13)$$

$$z_t = z_{p_j}, t = n + 1, \dots, 2n \text{ and } j = 1, 2, \dots, n, \text{ respectively,} \quad (2.14)$$

$$b_t = 0, t = 1, 2, \dots, n \quad (2.15)$$

$$b_t = 1, t = n + 1, \dots, 2n \quad (2.16)$$

\bar{E} exists if and only if

$$\alpha \leq \alpha_{\max} \quad (2.17)$$

where α_{\max} is an invariant defined on the interpolation data $\{z_t, b_t\}$ specified as in (2.12)–(2.16) above. Further, the mapping $E(s)$ given in Fig. 2 (where G is any arbitrary simply connected region) satisfying the condition given by (2.2) exists, if and only if

$$|\phi(1)| < \alpha_{\max}. \quad (2.18)$$

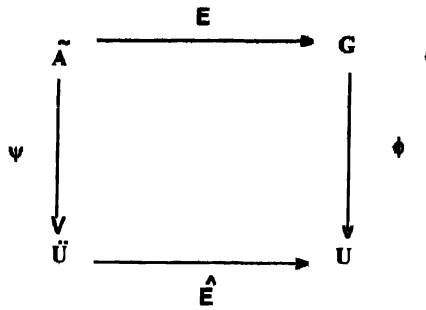
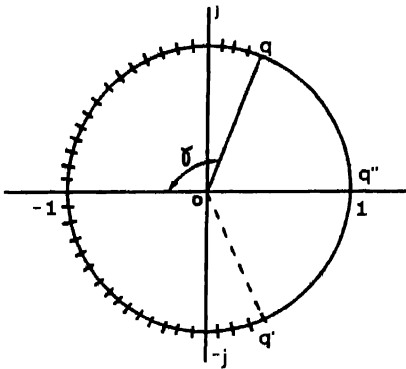


Fig. 2. Conformal mapping relationships.

Fig. 3. η -plane

III. TIME DELAY UNCERTAINTY

Consider the characteristic equation for the system of Fig. 1, viz.

$$1 + p_0(s) c(s) \exp(-T_d s) = 0, s \in \tilde{A}. \quad (3.1)$$

Using (2.1), (3.1) can be rearranged as

$$\eta(s) = \eta\{E(s)\} = \{1 - E(s)\}/E(s) = -\exp(-T_d s), s \in \tilde{A}. \quad (3.2)$$

The function η maps the E -plane into the η -plane. For $s = j\omega$, $\omega \geq 0$, the right-hand side (RHS) of (3.2) maps the $j\omega$ -axis in the s -plane onto the unit circle centre origin, in the η -plane as shown in Fig. 3.

Notice that

$$\eta(1) = 0 \quad (3.3)$$

and

$$\eta(0) = \infty. \quad (3.4)$$

Consider a simply connected region G' (containing the points 0 and 1) in the E -plane such that its image G'' (containing points ∞ and 0, respectively) under the mapping η corresponds to the η -plane minus the hatched arc as shown in Fig. 3. Using (3.2), if we define the mapping φ such that

$$\varphi\{\eta(s)\} = 1/\{1 + \eta(s)\} = E(s) \quad (3.5)$$

then G can be obtained by mapping G'' under φ . If G' can now be conformally mapped [6] in to the open unit disk U' by the mapping ϕ' such that

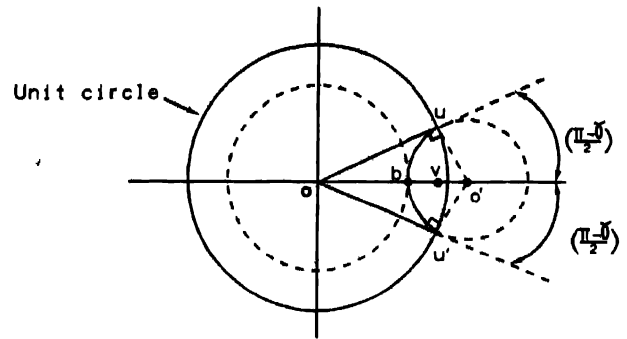
$$\phi'(\infty) = 0 \quad (3.6)$$

then the mapping $\phi: G \rightarrow U'$ of Fig. 2 will have been implemented (compare (2.4) with (3.4) and (3.6)) such that

$$\phi = \phi' \cdot \eta. \quad (3.7)$$

Let the mapping ϕ' be given by the function

$$\phi'(y) = \frac{\sqrt{\{(y-q)/(y-q')\} \exp(j\gamma)} - \exp(j\gamma/2)}{\sqrt{\{(y-q)/(y-q')\} \exp(j\gamma)} + \exp(-j\gamma/2)}$$

Fig. 4. ϕ' -plane.

$$\times [\exp\{j(\pi - \gamma)/2\}] \quad (3.8)$$

where

$$y = \eta\{E(s)\}$$

q and q' in the η -plane are conjugates of each other and

$$q = \exp\{j(\pi - \gamma)\} \quad (3.9)$$

γ is the angle as shown in Fig. 3. Also, from (3.3) and (3.7)

$$\phi(1) = \phi'(0). \quad (3.10)$$

The image of G' in Fig. 3 under the mapping ϕ' is as shown in Fig. 4. The hatched portion in the η -plane is mapped onto the unit circle in the ϕ' -plane. The arc $qq''q'$ in the η -plane is transformed into the arc $u'bu$ in the ϕ' -plane. Also, the origin in the η -plane is mapped on to the point v in the ϕ' -plane. Simple calculation shows that

$$\phi(1) = \phi'(0) = ov = \sin(\gamma/2), 0 < \gamma < \pi \quad (3.11)$$

and that

$$ob = \tan(\gamma/4). \quad (3.12)$$

Let

$$\tilde{E} = E \cdot v. \quad (3.13)$$

The function $\tilde{E}(s)$ can be computed from Fig. 2, (3.7) and (3.13) as

$$\tilde{E}(s) = \tilde{E}(s) \cdot v^{-1} = \frac{\{\tilde{E}(s) \sin(\gamma/2) - 1\} \{\tilde{E}(s)\}}{[\sin(\gamma/2) \{\tilde{E}(s)\}^2 - 1]} \quad (3.14)$$

and the corresponding controller $c(s)$ can be determined using (2.1) as

$$c(s) = \frac{[\{\tilde{E}(s) \sin(\gamma/2) - 1\} \{\tilde{E}(s)\}]}{p_0(s) [\{\tilde{E}(s)\} - \sin(\gamma/2)]} \quad (3.15)$$

The following result can now be stated.

Theorem 1. Consider the conformal mapping of Fig. 2. Let the simply connected region G be such that η maps G on to G'' where G'' corresponds to the η -plane minus the hatched arc on the unit circle defined by angle γ ($0 < \gamma < \pi$) as shown in Fig. 3 such that the mapping ϕ' is defined in terms of γ through (3.8) and (3.9). The mapping E satisfying the condition (2.2), then exists if and only if

$$\gamma < 2 \sin^{-1}(\alpha_{\max}). \quad (3.16)$$

The function $E(s)$ is given by (3.14).

Proof: From the result of [5], (2.18) follows. Since $\sin(\gamma/2)$ is positive for $0 < \gamma < \pi$, it then follows from (3.11) and (2.18) that

$$\sin(\gamma/2) < \alpha_{\max}.$$

Hence (3.16) follows. This ensures that $\tilde{E}(s)$ exists as per [5] based on the modified Nevanlinna-Pick theory. The procedure for finding $\tilde{E}(s)$ is well known [7], [8] through the formation of the Fejervary array corresponding to the mapping $\tilde{U} \rightarrow U'$. $\tilde{E}(s)$ is found from (3.13). Hence $E(s)$ is given by (3.14). Hence the result.

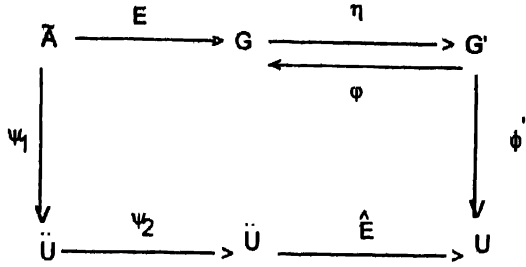


Fig. 5. Detailed conformal mappings.

IV. TIME DELAY MARGIN

Consider the transformation

$$v_1(s) = v = \{k(s - t)\}/(s + t), \quad 0 < k < 1, \quad t > 0 \quad (4.1)$$

such that

$$v_1: \tilde{A} \rightarrow \tilde{U}. \quad (4.2)$$

Further consider the transformation

$$v_2(v) = z = (v - k)/(1 - vk) \quad (4.3)$$

such that

$$v_2: \tilde{U} \rightarrow \tilde{U}. \quad (4.4)$$

Let

$$v = v_2 \circ v_1 \quad (4.5)$$

such that

$$v: \tilde{A} \rightarrow \tilde{U}. \quad (4.6)$$

The $j\omega$ -axis of the complex plane \tilde{C} together with the point at infinity is mapped by v_1 on to the circle (call it $c1$) centre origin and radius k which in turn is mapped by v_2 onto the circle (call it $c2$) on a diameter whose extreme points correspond to $\{-2k/(1+k^2)\}$ and the origin. Notice that

$$v(\infty) = v_2\{v_1(\infty)\} = 0. \quad (4.7)$$

Referring to Fig. 2, if the zero at infinity of E is of order $m' (=n-m)$ then so is the zero of \tilde{E} at the origin. Since \tilde{E} is analytic in \tilde{U} , it follows by Schwarz's lemma [9] that

$$\{|\tilde{E}(z)|\} < |z|^{m'}, \quad |z| < 1. \quad (4.8)$$

Fig. 5 shows the detailed mapping as discussed so far.

Representing an arbitrary point on the circle $c1$ as

$$v = v_1 = k \exp(j\theta), \quad 0 \leq \theta \leq 2\pi. \quad (4.9)$$

Consideration of the mapping v_1 will show that

$$\tan(\theta/2) = t/\omega. \quad (4.10)$$

Further, considering the mapping v_2 , it can be shown that

$$\begin{aligned} |z| &= |z_1| = f_1(\omega) \\ &= \frac{2kt}{\sqrt{(1-k^2)^2\omega^2 + (1+k^2)^2t^2}}, \quad \omega \geq 0 \end{aligned} \quad (4.11)$$

where

$$z_1 = v_2(v_1). \quad (4.12)$$

Define

$$\rho(\gamma) = \{\tan(\gamma/4)\}^{1/m'}. \quad (4.13)$$

Let ω be such that

$$f_1(\omega) < \rho(\gamma). \quad (4.14)$$

Substituting for $f_1(\omega)$ from (4.11), (4.14) can be expressed as

$$\omega > \omega_0 = \frac{t}{(1-k^2)} \sqrt{\frac{4k^2}{\rho^2(\gamma)} - (1+k^2)^2}. \quad (4.15)$$

We can now state the following result.

Theorem 2. Assume that the condition of Theorem 1 is met and that initially the nominal system is stable. Then the closed-loop system is stable under the time delay uncertainty provided T_d satisfies the inequality

$$T_d < \gamma/\omega_0 \quad (4.16)$$

where ω_0 is given by the RHS of (4.15).

Proof: Following generally the approach due to Walton and Marshall [3], since the control system of Fig. 1 is nominally stable, under time delay uncertainty, the closed-loop poles first need to cross the imaginary axis in the complex plane before the system becomes unstable. The zeros of the characteristic (3.1) are also necessarily solutions of

$$p_0(j\omega) p_0(-j\omega) e(j\omega) e(-j\omega) = 1, \quad \omega \geq 0 \quad (4.17)$$

since (3.1) is real. The number of ω 's (they are all finite) which correspond to the solution of (4.17) are finite in number. Let the largest of them be denoted by ω_m .

A necessary condition for the roots of the characteristic (3.1) to cross the imaginary axis is when the gain condition of (3.2) (equivalent to (3.1)) is satisfied for $s = j\omega$. The RHS of (3.2), when $s = j\omega$ corresponds to points on the unit circle in the η -plane. Since the points on the hatched portion of the unit circle are left out by design from the η -mapping, solutions to (3.2) can only exist corresponding to points on the arc $qq''q'$ in the η -plane (see Fig. 3). The mapping ϕ' maps the arc $qq''q'$ on to the arc $u'hu$ (see Fig. 4) as explained in Section III. The smallest distance from the origin to points on the arc $u'hu$ corresponds to the length of the line segment ob (see Fig. 4) which is given by (3.12).

Considering the mapping of the $j\omega$ -axis via the transformation $\{E \circ v\}$ (see Fig. 2) onto the open unit disk U' and using (4.8) through (4.12), it follows that

$$E(z_1) < |z_1|^{m'} = \{f_1(\omega)\}^{m'}. \quad (4.18)$$

Equation (4.14) then ensures that the mapping of the $j\omega$ -axis via the transformation $\{E \circ v\}$ on to the unit open disk U' (see Fig. 4) falls short of the arc $u'hu$, from magnitude considerations alone for all $\omega > \omega_0$ where ω_0 is given by the RHS of (4.15). Since the arc $u'hu$ is the image under the mapping ϕ' of the arc $qq''q'$ corresponding to feasible ω -solutions of (3.2), it follows that (3.2) cannot have a solution for $\omega > \omega_0$. It then follows that

$$\omega_0 \geq \omega_m.$$

Since the angle subtended by the circular arc between any point on the arc $qq''q'$ and the (-1) point in Fig. 3 at the origin, can be considered (see RHS of (3.2) as corresponding to the product $\{\omega T_d\}$, the smallest such angle corresponds to the point q , the angle being γ as shown in Fig. 3. Since no solution to (3.2) exists for $\omega > \omega_0$, the same is true for

$$T_d < \gamma/\omega_0$$

and hence the closed-loop stability is assured. Hence the result.

Remark 1: The analysis leading to Theorem 2 is based on the conformal mapping $v_1(s)$ (see (4.1)) for $0 < k < 1$. The strict inequality, viz., $k < 1$ is necessary to include the case of zeros of $p_0(s)$, lying on the $j\omega$ -axis and at infinity, being mapped to zero in our analysis. This is due to the well-known result [10] that the

Möbius transformation [represented by (4.3) in our case] can only map points on the circumference on the unit circle on to itself. A provision to the effect of $k < 1$ is also assumed while considering the boundary interpolation case in theorem 1.5 of [5]. Moreover, in our case, as will be shown in Theorem 3, a lower bound for T_{\max} is optimized in a certain sense over k , $0 < k < 1$.

Remark 2: The ratio γ/ω_0 represents a lower bound for T_{\max}

$$\gamma/\omega_0 \leq T_{\max}.$$

For a given β and k , the ratio (γ/ω_0) is maximized by choosing γ as large as possible subject to the constraint of inequality (3.16). Now it remains to optimize (maximize) the lower bound, in a certain sense, by choosing k appropriately. This is taken up next.

In the sequel, k is treated as a variable, $0 < k < 1$, and all the parameters, such as, α_{\max} , γ , ω_0 , ρ and others are all dependent on k .

Theorem 3 Consider the lower bound, viz., the ratio (γ/ω_0) for the time delay margin T_{\max} as given in Theorem 2. For $0 < k < 1$, a maximum exists for this ratio if

$$\rho(k) < (1/a)\{2k/(1+k^2)\}, \quad 0 < k < 1, a > 1 \quad (4.19)$$

and this maximum corresponds to the value of k such that the following conditions are satisfied

$$\begin{aligned} (d\gamma/dk) \rho \{1k^2 - (1+k^2)^2 \rho^2\} + (d\rho/dk) 4k^3 \\ = 4\gamma k \rho \{(1+k^2)/(1-k^2)\}(1-\rho^2), \quad 0 < k < 1 \end{aligned} \quad (4.20)$$

$$(1/\omega_0)(d^2\omega_0/dk^2) > (1/\gamma)(d^2\gamma/dk^2) \quad (4.21)$$

where

$$\begin{aligned} d^2\omega_0/dk^2 = H(k)\{B(k) + L(k)(d\rho/dk) \\ + D(k)(d\rho/dk)^2 + F(k)(d^2\rho/dk^2)\} \end{aligned} \quad (4.22)$$

$$H(k) = [4\beta/\{(1-k^2)^3\rho^6\}][(4k^2/\rho^2) - (1+k^2)^2]^{-1/2} \quad (4.23)$$

$$\begin{aligned} B(k) = -12k^3(1-k^2)^2\rho + 8k^3(3+k^2)\rho^2 \\ + 3k(1-k^2)^2\rho^3 \\ - (1+6k^2+36k^4+18k^6+3k^8)\rho^4 \\ + (1+k^2)^3(1+3k^2)\rho^5 \end{aligned} \quad (4.24)$$

$$\begin{aligned} L(k) = (1-k^2)\{12k^4(1-k^2) + \rho(4k^3)(1-5k^2) \\ - \rho^2(1+k^2)^2(1-k^2)(3k^2) \\ + \rho^3(1+k^2)k(1+7k^4)\} \end{aligned} \quad (4.25)$$

$$D(k) = -(1-k^2)^2(4k^4) \quad (4.26)$$

$$F(k) = -\{4k^2 - \rho^2(1+k^2)^2\}(1-k^2)^2k^2\rho \quad (4.27)$$

and for clarity, the argument k is dropped from $\rho(k)$ and written as ρ only.

Proof Using (4.19), the RHS of (4.15) can be written as

$$\omega_0 > \omega_1 = [\beta\{1/(1+k^2)\}/(1-k^2)]\{\sqrt{a^2-1}\}. \quad (4.28)$$

Since $\beta > 0$ and $a > 1$ are constants, it follows that

$$\omega_1 \rightarrow \beta\sqrt{a^2-1} > 0, \quad k \rightarrow 0. \quad (4.29)$$

Also, for k in the range $0 < k < 1$, since the interpolation data $\{z_l, b_l\}$, $l = n+1, \dots, 2n$, $n \neq 0$ (specified in (2.14) and (2.16)) is confined to the region enclosed by the circle ϵ_2 in \tilde{t}^* (the image of \tilde{A} under the mapping ψ), it follows by Schwarz's Lemma (applied to the analytical map \tilde{E} given by (2.11)) that necessarily

$$\alpha_{\max} < \{2k/(1+k^2)\}^{m'}, \quad m' \geq 1. \quad (4.30)$$

Due to (3.16) and (4.30)

$$\alpha_{\max} \rightarrow 0 \text{ and } \gamma \rightarrow 0, \text{ as } k \rightarrow 0. \quad (4.31)$$

Hence using (4.28), (4.29) and (4.31)

$$(\gamma/\omega_0) < (\gamma/\omega_1) \rightarrow 0, \text{ as } k \rightarrow 0. \quad (4.32)$$

Also, as $k \rightarrow 1$, $\omega_1 \rightarrow \infty$. With $0 < \gamma < \pi$

$$(\gamma/\omega_0) < (\gamma/\omega_1) \rightarrow 0, \text{ as } k \rightarrow 1. \quad (4.33)$$

Notice that ω_0 given by RHS of (4.15) is such that $\omega_0 > 0$, $0 < k < 1$ since $\rho < 1$. That $\rho < 1$ can be seen from the inequalities (4.34) and (4.35) below

$$\begin{aligned} \tan(\gamma/4) < \sin(\gamma/2) < \alpha_{\max} \leq \{2k/(1+k^2)\}^{m'}, \\ m' \geq 1, 0 < \gamma < \pi \end{aligned} \quad (4.34)$$

$$\begin{aligned} \rho = \{\tan(\gamma/4)\}^{1/m'} < 2k/(1+k^2) < 1, \\ 0 < k < 1. \end{aligned} \quad (4.35)$$

With $\gamma > 0$, then the ratio $(\gamma/\omega_0) > 0$. (4.32) and (4.33) thus reveal that at least a maximum value for the ratio (γ/ω_0) exists in the range $0 < k < 1$.

Differentiating the ratio (γ/ω_0) with respect to k (with ω_0 given by RHS of (4.15)) and equating it to zero yields, after simplification, (4.20). To ensure that (4.20) corresponds to a maximum condition, one needs to check whether

$$d^2(\gamma/\omega_0)/dk^2 < 0. \quad (4.36)$$

Direct differentiation of the RHS of inequality (4.16) with respect to k yields (4.21). Equation (4.22)–(4.27) are obtained by some simplification of the direct differentiation of ω_0 given by the RHS of inequality (4.15) with respect to k . Hence the result of Theorem 3.

Remark 3 That the condition given in the inequality (4.19) is only a mild one can be seen by comparing the constraint on $\rho(k)$ due to the inequalities (4.19) and (4.35).

V. EXAMPLE

Let

$$\begin{aligned} p_0(s) &= 1/(s-2) \\ \beta &= 1, \quad m' = 1. \end{aligned}$$

Assuming k as a variable, using (4.1) and (4.3)

$$v_1(2) = v_1 = k/3$$

$$v_1(\infty) = v_2 = k$$

and

$$z_1 = v_2(v_1) = -2k/(3-k^2)$$

$$z_2 = v_2(v_2) = 0.$$

For the interpolation data of $(z_1, 1)$ and $(z_2, 0)$ for the mapping $\tilde{t}^* \rightarrow \tilde{t}$

$$\alpha_{\max} = 2k/(3-k^2).$$

Since $\sin(\gamma/2)$ can be arbitrarily close to α_{\max} , put

$$\sin(\gamma/2) = \alpha_{\max}.$$

$$\gamma = 2\sin^{-1}\{2k/(3-k^2)\}$$

$$\begin{aligned} \rho &= \tan(\gamma/4) = \{1 - \sqrt{1 - \alpha_{\max}^2}\}/\alpha_{\max} \\ &= \{(3-k^2) - \sqrt{(9-k^2)(1-k^2)}\}/(2k) \end{aligned}$$

$$d\rho/dk = (3+k^2)\rho/\{k\sqrt{(9-k^2)(1-k^2)}\}$$

$$d\gamma/dk = 4(3+k^2)/\{(3-k^2)\sqrt{(9-k^2)(1-k^2)}\}.$$

Equation (4.20) is satisfied in the range $0 < k < 1$, for $k = 0.781$ only. Also, using (4.22)–(4.27), (4.21) has been verified to be true at $k = 0.781$. The ratio γ/ω_0 is maximized for $k = 0.781$ and a lower bound for T_d is obtained as

$$T_{d\max} \geq \gamma/\omega_0 = 1.424519519/9.942663337 \\ = 0.143273434.$$

VI. CONCLUSION

Given a plant transfer function with time delay uncertainty, the question arises as to the maximum limiting time delay below which a controller exists such that the closed-loop system remains stable. Since no restriction is placed on the controller, it is expected that the maximum limiting time delay should be a function of the plant transfer function only. This paper has provided a solution to this problem in terms of a lower bound for the limiting time delay. The result will be helpful in determining the time delay margin below which the closed-loop stability of a given plant is guaranteed with the existence of a stabilising controller which can be determined in a straight forward way.

REFERENCES

- [1] K. Walton and J. E. Marshall, "Direct method for TDS stability analysis," *IEE Proc*, vol. 134, Pt D, no. 2, 1987, pp. 101–107.
- [2] El Sakkary, "Estimating robust dead time for closed loop stability," *IEEE Trans Automat Contr*, vol. 35, no. 2, pp. 209–210, 1990.
- [3] G. Zames and B. A. Francis, "Feedback, minimax sensitivity and optimum robustness," *IEEE Trans Automat Contr*, vol. 28, no. 5, pp. 585–601, 1983.
- [4] D. C. Youla, J. Bongiorno, and Y. Lu, "Single loop feedback stabilization of multivariable dynamic plants," *Automatica*, vol. 10, pp. 159–173, 1974.
- [5] P. P. Khargonekar and A. Tannenbaum, "Non-euclidean metrics and robust stabilization of systems with parameter uncertainty," *IEEE Trans Automat Contr*, vol. AC-30, no. 10, pp. 1005–1013, 1985.
- [6] L. Ahlfors, *Complex Analysis*. New York: McGraw-Hill, 1979.
- [7] H. Kimura, "Robust stabilizability for a class of transfer functions," *IEEE Trans Automat Contr*, vol. AC-29, no. 9, pp. 788–793, 1984.
- [8] N. I. Akhizer, "The classical moment problem and some related questions in analysis," Kemmer, Oliver and Boyd, Ed., 1965.
- [9] W. Rubin, *Real and Complex Analysis*. New York: McGraw-Hill, 1986.
- [10] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum, *Feedback Control Theory*. New York: Macmillan, 1992.

A New Reduction Technique for a Class of Singularly Perturbed Optimal Control Problems

Vladimir Gaitsgory

Abstract—We consider a class of singularly perturbed optimal control problems which may not be approximated by the reduced problems constructed via the formal replacement of the fast variables by the states of equilibrium of the "fast" subsystem considered with "frozen" slow variables and controls. We construct a reduced optimal control problem which provides a true approximation for the problems under consideration and write down necessary and sufficient optimality conditions for this reduced problem.

Manuscript received February 28, 1994.

The author is with the School of Mathematics, University of South Australia, The Levels, Pooraka, South Australia 5095.

IEEE Log Number 9408779

I. INTRODUCTION

Singularly perturbed optimal control problems have been intensively studied in the literature (see the overviews in [1], [7]–[9] and also more recent works [5], [6], [10]–[12]). As follows from the quoted overviews, a common approach to singularly perturbed optimal control problems is based on their replacement by some reduced order problems obtained via equating of the singular perturbation parameter to zero.

Recently in [5], [6], [11], and [12], it was shown that there are classes of problems which do not allow such an approach. That is, equating the parameter to zero leads, of course, to a simpler problem but the solution of this problem does not approximate the solution of the original singularly perturbed one with an arbitrarily small but nonzero value of the parameter.

In [11] and [12], it was noticed that such a situation may appear in H^∞ -optimal control problems. In [5] and [6], it was established that it may appear in the standard optimal control setting if the system is nonlinear and/or the cost function is nonconvex.

Our consideration in this paper is connected with results in [5] about an approximation of singularly perturbed optimal control problems by problems of optimization of some differential inclusions. In contrast to [5], where the main goal was to construct an approximating problem, in this paper we are interested in how this problem can be solved. We study here an interesting special case where the approximating optimization problem allows a simple representation as a "classic" optimal control problem which we call an extended reduced problem. We construct necessary and sufficient optimality conditions for this problem and describe how it is related to the reduced problem obtained in a traditional way (that is, via equating the small parameter to zero).

II. EXTENDED REDUCED PROBLEM

We shall restrict ourselves to consideration of a class of singularly perturbed (SP) optimal control problems written in the form

$$\text{minimize } G(z(1)) \quad (1)$$

subject to

$$\dot{z}(t) = Q(z(t)) + A(z(t))g(y(t), u(t)), \quad (2)$$

$$\epsilon \dot{y}(t) = f(y(t), u(t)), \quad y(0) = y \quad (3)$$

$$u(t) \in U, \quad \forall t \in [0, 1]. \quad (4)$$

Here ϵ is a small parameter. $Q(\cdot)$, $g(\cdot)$, $f(\cdot)$ are vector functions taking their values in R^k , R^l and R^n , respectively, and $A(\cdot)$ is a $k \times l$ matrix function. The functions $G(\cdot)$, $Q(\cdot)$, and $A(\cdot)$ are supposed to satisfy local Lipschitz conditions in z and the functions $g(y, u)$, $f(y, u)$ are supposed to be continuous and satisfy local Lipschitz conditions in y with a constant which does not depend on $u \in U$. The minimization is executed on the set of admissible controls which is defined as the set of measurable functions $u(t)$ satisfying the inclusion (4), where U is a compact subset of R^m .

We shall suppose that subsystem (3) being written in the "stretched" time scale $\tau = t\epsilon^{-1}$

$$\dot{y}(\tau) = f(y(\tau), u(\tau)) \quad (5)$$

satisfies the following two assumptions.

Assumption 1: There exists a compact set $P \subset R^n$ and a function $\eta(\tau)$

$$\lim_{\tau \rightarrow \infty} \eta(\tau) = 0, \quad \int_0^\infty \eta(\tau) d\tau < \infty$$

such that for any $y' \in P$, $i = 1, 2$ and for any admissible control $u(\tau)$: $\|y(\tau, u(\cdot), y^1) - y(\tau, u(\cdot), y^2)\| \leq \eta(\tau) \|y^1 - y^2\|$, where $y(\tau, u(\cdot), y)$ stands for the solution of (5) obtained with the control $u(\tau)$ and with the initial values $y(0) = y$.

Assumption 2: There exists a compact set $\Omega \subset P$ such that for any admissible control $u(\tau)$ and for any initial values $y \in \Omega$, the solution of (5) does not leave P .

Along with these two we also need an assumption about the trajectories of the SP system (2)–(4).

Assumption 3: There exist compact sets D and W , $D \subset W \subset R^n$, such that for any admissible control $u(t)$ and for any initial values $(z, y) \in D \times \Omega$ the trajectories of (2)–(4) do not leave $W \times P$.

Under Assumptions 1–3 from Theorems 3.2 and 4.1 in [5] it follows that the z -components of the trajectories of (2)–(4) are approximated by the solutions of the differential inclusion with the right-hand side defined for all $z \in W$ as the closure $\text{cl}\{Q(z) + A(z)V_p\}$, where the set $V_p \subset R^l$ is the union

$$V_p = \bigcup_{T>0} \bigcup_{u(\cdot)} \left\{ T^{-1} \int_0^T g(y(\tau), u(\tau)) d\tau \right\} \quad (6)$$

taken over all admissible controls $u(\tau)$ defined on the interval $[0, T]$, and also over all $T > 0$, with $y(\tau)$ being the solution of (5) obtained with the control $u(\tau)$ and satisfying the periodicity condition

$$y(0) = y(T) \in P. \quad (7)$$

Since $\text{cl}\{Q(z) + A(z)V_p\} = Q(z) + A(z)\text{cl}\{V_p\}$ and since the closure $\text{cl}\{V_p\}$, which we denote as \bar{V} , is compact and convex [5, Theorems 3.1 and 3.2], the set of the solutions of the above mentioned differential inclusion coincides with the set of trajectories of the system

$$\dot{z}(t) = Q(z(t)) + A(z(t))v(t) \quad z(0) = z \quad (8)$$

with the admissible controls $v(t)$ being defined as the measurable functions satisfying the inclusion

$$v(t) \in \bar{V}, \quad \forall t \in [0, 1]. \quad (9)$$

The problem of minimization of the cost function (1) on the trajectories of system (8) with the admissible controls defined as above will be called the extended reduced (ER) problem.

An application of Theorem 4.1 of [5] leads now to the following statements about the connections between the SP and ER problems: There exists a function $\mu(\epsilon)$ tending to zero as ϵ tends to zero such that:

- 1) The optimal value function $W^*(s, z, y)$ of the SP problem and the optimal value function $W^*(s, z)$ of the ER problem (that is, the optimal values of the cost achieved in the SP and ER problems subject to that the motions begin from the moment $s \in [0, 1]$ with the initial values $(z(s), y(s)) = (z, y)$ and $z(s) = z$, respectively) satisfy the inequality

$$|W^*(s, z, y) - W^*(s, z)| \leq \mu(\epsilon),$$

$$\forall (s, z, y) \in [0, 1] \times D \times \Omega.$$

- 2) Let $\delta \geq 0$ be an arbitrary number and $v^\delta(t)$ be an admissible control in the ER problem such that

$$G(z^\delta(1)) \leq W^*(0, z) + \delta \quad (10)$$

where $z^\delta(t)$ is the trajectory of (8) obtained with the control $v^\delta(t)$ and with initial values $z^\delta(0) = z \in D$, that is, $v^\delta(t)$ is δ -suboptimal in the ER problem. Then using this control one may find a control $u^\delta(t)$ which is asymptotically δ -suboptimal in the SP problem. That is, it satisfies (4) and when using in (2), (3) with the initial values $(z^\delta(0), y^\delta(0)) = (z, y)$, $y \in \Omega$ it generates the trajectory $(z^\delta(t), y^\delta(t))$ satisfying the inequality

$$G(z^\delta(1)) \leq W^*(0, z, y) + \delta + O(\mu(\epsilon)).$$

Notice that in a general case the ER problem is not equivalent to the reduced (R) problem obtained via equating ϵ to zero in (1)–(4). The latter, as can be easily shown on the basis of Filipov's Lemma [3], is equivalent to the problem of minimization of cost function (1) on the trajectories of system (8) with admissible controls $v(t)$ being measurable functions satisfying the inclusion

$$v(t) \in V_{\epsilon} \quad (11)$$

where $V_{\epsilon} = \bigcup_{u \in U} \{g(v(u), u)\}$ and $v(u)$ is the root of the equation: $f(y, u) = 0$. Since, by Theorem 3.2 [5], $\text{conv} V_{\epsilon} \subset \bar{V}$ and this inclusion can be proper, all the trajectories which are admissible in the R problem are also admissible in the ER problem but not vice versa.

III. NECESSARY AND SUFFICIENT CONDITIONS OF OPTIMALITY FOR THE ER PROBLEM

Let us introduce the following periodic optimization PO problem

$$\sup_{u(\cdot)} \left\{ T^{-1} \int_0^T \lambda^T A(z) g(y(\tau), u(\tau)) d\tau \right\} = h(z, \lambda). \quad (12)$$

Here sup is sought over the length T of the time interval, over the admissible controls $u(\tau) \in U$ defined on $[0, T]$ and over the corresponding solutions $y(\tau)$ of system (5) satisfying periodicity conditions (7). The optimal value of the problem depends on the values of the parameters z , λ and the function defining this dependence is denoted as $h(z, \lambda)$.

Let $(v^0(t), z^0(t))$ be an optimal control and the corresponding trajectory of the ER problem. Since, by definition, $v^0(t) \in \bar{V}$, $\forall t \in [0, 1]$ and \bar{V} is the closure of the set V_p introduced in (6), corresponding to each $t \in [0, 1]$ there exist a sequence of positive numbers T_i^t , a sequence of controls $u_i^t(\tau) \in U$ defined on the intervals $[0, T_i^t]$ and a sequence of the solutions $y_i^t(\tau)$ of system (5) satisfying the periodicity conditions: $y_i^t(0) = y_i^t(T_i^t) \in P$, $i = 1, 2, \dots$, such that

$$v^0(t) = \lim_{i \rightarrow \infty} (T_i^t)^{-1} \int_0^{T_i^t} g(y_i^t(\tau), u_i^t(\tau)) d\tau. \quad (13)$$

Theorem 1 Let Assumptions 1–3 hold and let $G(z)$, $Q(\cdot)$, and $A(\cdot)$ have continuous partial derivatives with respect to the components of z . Let $\lambda(t)$ be the solution of the system

$$\dot{\lambda}(t) = -F'_z(z^0(t), v^0(t))\lambda(t), \quad \lambda(1) = -G'(z^0(1)) \quad (14)$$

where

$$\begin{aligned} F(z, v) &= Q(z) + A(z)v, \\ F'_z(z, v) &= \left\{ \frac{\partial F_i(z, v)}{\partial z_j} \right\}, \\ G'_z(z) &= \left\{ \frac{\partial G_i(z)}{\partial z_j} \right\} \end{aligned} \quad (15)$$

$F_i(z, v)$, $i = 1, \dots, k$ are the components of the vector function $F(z, v)$. Then for almost all $t \in [0, 1]$

$$\begin{aligned} \lim_{i \rightarrow \infty} (T_i^t)^{-1} \int_0^{T_i^t} \lambda^T(t) A(z^0(t)) g(y_i^t(\tau), u_i^t(\tau)) d\tau \\ = h(z^0(t), \lambda(t)). \end{aligned} \quad (16)$$

That is, the sequence of triplets $\{T^i, u^i(\tau), y^i(\tau)\}$ defined as that satisfying (13) is maximizing in PO problem (12) with $z = z^0(t)$ and $\lambda = \lambda(t)$.

Proof of the theorem is obtained via the direct application of Pontriagin's maximum principle to the ER problem.

Let us note that if for all t from some subset ω of $[0, 1]$ there exist $T^i > 0$, a control $u^i(\tau) \in U$, $\forall \tau \in [0, T^i]$, and the solution $y^i(\tau)$ of system (5) satisfying periodicity conditions: $y^i(0) = y^i(T^i) \in P$ such that

$$v^0(t) = (T^i)^{-1} \int_0^{T^i} g(y^i(\tau), u^i(\tau)) d\tau \quad (17)$$

that is, $v^0(t) \in V_p$, $\forall t \in \omega$, then for almost all such t the triple $\{T^i, u^i(\tau), y^i(\tau)\}$ is the solution of PO problem (12) with $z = z^0(t)$, $\lambda = \lambda(t)$.

Let us denote by $\{T^{-\lambda}, u^{-\lambda}(\tau), y^{-\lambda}(\tau)\}$ a solution of PO problem (12) with given z and λ and define the function $v(z, \lambda)$ as

$$(T^{-\lambda})^{-1} \int_0^{T^{-\lambda}} g(y^{-\lambda}(\tau), u^{-\lambda}(\tau)) d\tau. \quad (18)$$

Assuming that (17) takes place for all $t \in [0, 1]$, one may come to the conclusion that the optimal control $v^0(t)$ can be presented in the form

$$v^0(t) = v(z^0(t), \lambda(t)) \quad (19)$$

where $z^0(t)$, $\lambda(t)$ are determined as a solution of the following boundary-value problem

$$\begin{aligned} \dot{z}^0(t) &= Q(z^0(t)) + A(z^0(t))v(z^0(t), \lambda(t)), \quad z^0(0) = z \\ \dot{\lambda}(t) &= -F'(z^0(t), v(z^0(t), \lambda(t)))\lambda(t), \\ \lambda(1) &= -G'(z^0(1)). \end{aligned}$$

Notice that the verification of (17) and the construction of the family of the solutions of (12) depending on z and λ seem to be very problematic in the general case. The described procedure, however, maybe also applied without such a verification and with the use of a family of approximate solutions of (12) instead of the family of precise ones. In Section IV below we shall consider a special case when an estimation of "suboptimality" of the control (19) via the estimation of the approximation of the solutions of (12) is possible.

Sufficient optimality conditions for the ER problem can be constructed on the basis of the Hamilton-Jacobi-Bellman (HJB) equation written for this problem in the form

$$W'_z(s, z) + \min_{v \in V} \{ (W'_z(s, z))^T (Q(z) + A(z)v) \} = 0. \quad (20)$$

Using again the fact that \bar{V} is the closure of V_p and also the definition of $h(z, \lambda)$ as the optimal value of the PO problem (12), one may rewrite (20) in the following form

$$W'_z(s, z) + (W'_z(s, z))^T Q(z) - h(z, -W'_z(s, z)) = 0. \quad (21)$$

HJB equation (21) provides a characterization of the optimal value function of the ER problem (and, thus, an approximate characterization of the optimal value function of the SP problem). Namely, if the former is differentiable at some point $(s, z) \in R^{k+1}$, then it satisfies (29) at this point [4, Chapter IV, Theorem 4.1]. On the other hand, given some admissible trajectory $z^0(t)$ of the ER problem (that is, a trajectory of (8) obtained with some admissible control $v^0(t)$), its optimality can be verified on the basis of the following sufficient optimality conditions.

Theorem 2: If under Assumptions 1-3 there exists a function $W(s, z)$ defined and having continuous partial derivatives in some δ -tube, Γ_δ , around $z^0(t)$

$$\Gamma_\delta = \{(t, z) \mid t \in [0, 1], \|z - z^0(t)\| < \delta\}, \quad \delta > 0$$

and also satisfying there (21) with the following boundary conditions

$$\begin{aligned} W(1, z) &= G(z), \quad \forall z \in \{z^1 \mid \|z^1 - z^0(1)\| < \delta\}; \\ W(0, z^0(0)) &\geq G(z^0(0)) \end{aligned}$$

then $(v^0(t), z^0(t))$ is a local solution of the ER problem, that is, for all admissible trajectories $z(t)$ such that $(t, z(t)) \in \Gamma_\delta$

$$G(z^0(1)) \leq G(z(1)).$$

Proof: The theorem is a reformulation of Proposition 3.7.2 in [2] and, thus, it is implied by the proof of this proposition.

In a general case the optimal value function of the ER problem is not differentiable at all points, although it can be shown to satisfy local Lipschitz conditions and, thus, it is differentiable almost everywhere [4, Chapter IV, Theorem 4.2]. Similarly to Theorem 2, sufficient (and also necessary) conditions for the admissible trajectory of the ER problem to be locally optimal can be constructed on the basis of some generalized version of the GJB equation introduced in [2]. This is

$$\min \{ \alpha + \beta^T Q(z) - h(z, -\beta), \quad (\alpha, \beta) \in \partial W(s, z) \}$$

where $\partial W(s, z)$ is the generalized Clarke gradient. The formulation and the proof of the corresponding statement can be adapted from [2, Theorem 3.7.6]. \square

IV. ASYMPTOTIC LINEARIZATION OF THE SP PROBLEM

Let us consider a class of SP problems which are linear in the slow variables. That is, assume that cost function (1) and slow subsystem (2) are written as

$$\inf_{u \in U} \int_0^1 z^T(t) dt \quad (22)$$

and

$$\dot{z}(t) = Qz(t) + Ag(y(t), u(t)), \quad z(0) = z \quad (23)$$

respectively, where $Q(k \times k)$ and $A(k \times l)$ are constant matrices. In this case the system (8) becomes linear

$$\dot{z}(t) = Qz(t) + Av(t), \quad z(0) = z \quad (24)$$

and, thus, by Theorem 1, nonlinear SP problem (22), (23), (3), (4) is approximated by linear ER problem (22), (24), (9).

The linearity of the ER problem makes Pontriagin's maximum principle not only a necessary but also a sufficient condition of optimality in this problem. More than that, the following statement is known to be true.

Let $\lambda(t)$ be defined as the solution of system (14) which, by (22) and (24) takes the form

$$\dot{\lambda}(t) = -Q^T \lambda(t), \quad \lambda(1) = -c \Leftrightarrow \lambda(t) = -e^{Q^T(1-t)} c \quad (25)$$

and let an admissible control $v^\delta(t)$ satisfies the inequality

$$\lambda^T(t)Av^\delta(t) \geq h(\lambda(t)) - \eta(t), \quad \eta(t) \geq 0$$

for almost all $t \in [0, 1]$. Here

$$h(\lambda) = \max \{ \lambda^T A v \mid v \in \bar{V} \} \quad (26)$$

and $\eta(t)$ is some integrable function.

Then the trajectory $z^\delta(t)$ of (24) obtained when using this control satisfies (10) with $\delta = \int_0^1 \eta(t) dt$. That is, $v^\delta(t)$ is δ -suboptimal in the ER problem.

The validity of this statement follows from the fact that the value of the cost function (22) obtained when applying the control $v^\delta(t)$ on the interval $[s, 1]$ with the initial values $z(s) = z$ is equal to

$$c^T z^\delta(1) = -\lambda^T(s)z - \int_s^1 \lambda^T(t)A v^\delta(t)dt$$

whereas the optimal cost function value is equal to

$$W(s, z) = -\lambda^T(s)z - \int_s^1 h(\lambda(t))dt. \quad (27)$$

Notice that

$$\max\{\lambda^T v \mid v \in \bar{V}\} = \sup\{\lambda^T v \mid v \in V_p\}$$

and, thus, $h(\lambda)$ coincides with the optimal value of the PO problem (12), with the dependence on z disappearing since the matrix $A(z)$ is supposed to be constant. This allows us to reformulate the statement above as follows.

Let T^λ be a positive number, $u^\lambda(\tau) \in U$ be a control defined on $[0, T^\lambda]$ and $y^\lambda(\tau)$ be the solution of (5) corresponding to this control satisfying the periodicity condition: $y^\lambda(0) = y^\lambda(T^\lambda) \in P$ such that

$$\begin{aligned} (T^\lambda)^{-1} \int_0^{T^\lambda} \lambda^T A g(y^\lambda(\tau), u^\lambda(\tau)) d\tau \\ \geq h(\lambda) - \eta(\lambda), \quad \eta(\lambda) \geq 0 \end{aligned}$$

and let

$$v(\lambda) = (T^\lambda)^{-1} \int_0^{T^\lambda} g(y^\lambda(\tau), u^\lambda(\tau)) d\tau.$$

Suppose that the family of triplets $\{T^\lambda, u^\lambda(\tau), y^\lambda(\tau)\}$ and, thus, also $v(\lambda)$ are defined for all $\lambda \in \Lambda$, with Λ containing $\lambda(t), \forall t \in [0, 1]$. Then the control $v^\delta(t) = v(\lambda(t))$ is δ -suboptimal in the ER problem with $\delta = \int_0^1 \eta(\lambda(t))dt$.

V. CONNECTIONS BETWEEN THE ER AND R PROBLEMS

As we have already noticed, the sets of controls and trajectories admissible in the ER problem are in the general case more rich than those in the R one and, hence, the optimal value functions of these two problems, $W(s, z)$ and, respectively, $F(s, z)$, satisfy the inequality

$$F(s, z) \geq W(s, z) \quad \forall (s, z) \in [0, 1] \times R^k.$$

In case the SP problem is linear in z , that is (22), (23) are valid, the difference between $F(s, z)$ and $W(s, z)$ allows an explicit representation.

Let us define $h_{st}(\lambda)$ as the optimal value of the steady state optimization SSO problem corresponding to PO problem (12)

$$\begin{aligned} h_{st}(\lambda) &= \max\{\lambda^T A g(v(u), u) \mid u \in U\} \\ &= \max\{\lambda^T A g(y, u) \mid g(y, u) = 0, u \in U, y \in P\} \end{aligned}$$

where again, as above, the dependence of the optimal value on z disappears since A is not dependent on z .

Theorem 3: Suppose that Assumptions 1-3 and (22), (23) are true. Then

$$F(s, z) - W(s, z) = \int_s^1 (h(\lambda(t)) - h_{st}(\lambda(t))) dt \quad (28)$$

where $\lambda(t)$ is defined in (25).

Proof: In the case under consideration the R problem takes the form (22), (24), (11). That is, it is linear and, hence, its optimal value function is given by the expression

$$F(s, z) = -\lambda^T(s)z - \int_s^1 \max\{\lambda^T(t)A v \mid v \in V_{st}\} dt.$$

Comparing this with (27) and taking into account that

$$\max\{\lambda^T A v \mid v \in V_{st}\} = h_{st}(\lambda)$$

one obtains (28). \square

Corollary: Under the assumptions of Theorem 3 the optimal value function of the SP problem converges to the optimal value function of the R problem

$$\lim_{\delta \rightarrow 0} W(s, z, y) = F(s, z) \quad \forall (s, z, y) \in [0, 1] \times D \times \Omega \quad (29)$$

if and only if for almost all $t \in [0, 1]$

$$h(\lambda(t)) = h_{st}(\lambda(t)). \quad (30)$$

Notice that in a general case $h(\lambda) \geq h_{st}(\lambda)$ and a number of tests allowing one to verify whether this inequality is the strict one or it takes the form of the equality were developed in periodic optimization theory (see references in [5], [6]). Notice also that if this inequality takes the form of the equality for all λ (that is, (30) is true for all λ and not only for $\lambda = \lambda(t)$), then as it follows from Theorems 3.2 and 4.2 in [5], the convergence (29) can be guaranteed for SP problems of a much more general structure than that we are dealing with in this paper.

VI. CONCLUSION

We have proposed a new method to deal with the class of SP problems described. The method consists of an approximation of the solution of the SP problem by the solution of the ER problem and of characterization of the latter by necessary and sufficient optimality conditions. For SP problems linear in the slow variables the method allows the complete answer to the question about the efficiency of the traditional reduction technique.

REFERENCES

- [1] A. Bensoussan, *Perturbation Methods in Optimal Control Problems*. New York: Wiley, 1989.
- [2] F. H. Clarke, *Optimization and Nonsmooth Analysis*. New York: Wiley, 1983.
- [3] A. F. Filippov, "To some questions of the theory of optimal control," *Vestnik Moskovskogo Universiteta, Seria Matem.*, vol. 3, pp. 16-26, 1967 (in Russian).
- [4] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*. Berlin-Heidelberg-New York: Springer-Verlag, 1975.
- [5] V. Gaitsgory, "Suboptimization of singularly perturbed control systems," *SIAM J. Contr. Optim.*, vol. 30, no. 5, pp. 1228-1249, 1992.
- [6] —, "Suboptimal control of singularly perturbed systems and periodic optimization," *IEEE Trans. Automat. Contr.*, vol. 38, no. 6, pp. 888-903, 1993.
- [7] P. V. Kokotovic, "Applications of singular perturbations techniques to control problems," *SIAM Review*, vol. 26, pp. 501-550, 1984.
- [8] P. V. Kokotovic, H. Khalil, and J. O'Reilly, *Singular Perturbations in Control Analysis and Design*. New York: Academic, 1986.
- [9] P. V. Kokotovic, R. E. O'Malley, and P. Sannuti, "Singular perturbations and order reduction in control theory," *Automatica*, vol. 12, pp. 123-132, 1976.
- [10] J. Lehoczky, S. P. Sethi, H. M. Soner, and M. I. Taksar, "An asymptotic analysis of hierarchical control of manufacturing systems under uncertainty," *Mathematics Op. Res.*, vol. 16, no. 3, 1991.
- [11] Z. Pan and T. Basar, " H^∞ -optimal control for singularly perturbed systems. Part 1: Perfect state measurement," *Automatica*, vol. 29, no. 2, pp. 401-423, 1993.
- [12] —, " H^∞ -optimal control for singularly perturbed systems. Part 2: Imperfect state measurements," *IEEE Trans. Automat. Contr.*, vol. 39, no. 2, pp. 280-299, 1994.

Pole Assignment with Robust Stability

Mark E. Halpern, Robin J. Evans, and Robin D. Hill

Abstract—This paper uses convex analysis for the pole assignment design of discrete-time SISO systems incorporating robust stability against norm bounded parametric perturbations in the plant transfer function. The method involves designing an overparameterized pole assignment controller for the nominal plant with the overparameterization chosen to reduce the size of changes in the closed-loop characteristic equation which result from plant perturbations. Sufficiency bounds on the l_p norm of the perturbation guaranteeing stability are obtained.

I. INTRODUCTION

For most real control problems, the behavior of the plant does not match exactly that of an assumed finite order linear time-invariant dynamic model. This is because the true plant usually has complicated dynamics which may be nonlinear and of high order.

This motivates the design of robust controllers, that is compensators which achieve acceptable performance in some sense for a range of plants. This then can allow for uncertainties in the plant model. Uncertainties are usually considered to fit into two broad categories, namely structured and unstructured uncertainties depending on the specificity of the knowledge of the perturbations with respect to the plant model structure. For example, uncertainties in the frequency response of a plant are unstructured because they relate to the plant as a whole. On the other hand, uncertainties in a parametric description of a plant are structured. The two main aspects of robustness are stability robustness and performance robustness. In this work we focus on design for stability robustness in the presence of structured parametric uncertainty in the plant.

Previous related work is described in [1]–[9]. Dahleh [1] used the small gain theorem to obtain necessary and sufficient conditions for closed-loop stability in the presence of l_∞ induced norm bounded coprime factor perturbations to the plant. In [2], Soh and Evans applied Kharitonov results [3] to characterize allowable perturbations of an interval plant in closed loop with a pole placement compensator. Kharitonov-type extreme point results do not hold, however, for a stability region which is a circle at the origin of the complex plane so they are not easily used for testing discrete-time stability so a bilinear transform was used in [2].

An approach which handles general stability regions and, thus allows the treatment of discrete-time systems, involves treating the vector of coefficients of an n th order closed-loop characteristic equation as a point in n -dimensional parameter space. Fam and Meditch [4] have determined the stability boundary for discrete-time systems in this space of coefficients and have shown it to be bounded by three hypersurfaces. Soh *et al.* [5] have found the largest hypersphere centered at a specified stable point and containing the coefficients of only stable polynomials.

Using this type of framework, Biernacki *et al.* [6], and Keel *et al.* [7] have proposed continuous-time compensator synthesis approaches for uncertain multivariable plants. They regard the compensator as a

mapping from plant parameter space to closed-loop coefficient space and design the compensator to restrict the closed-loop coefficients to the interior of a stable region in this space. The compensator design involves a nonconvex numerical optimization.

Berger [8], [9] has proposed an approach for the synthesis of robust control systems using overparameterized pole assignment controllers. In that work, the overparameterization is used to minimize the variance of the closed-loop characteristic polynomial due to plant parameter uncertainties described by the covariance matrix of the plant parameter estimates as obtained by a recursive least squares estimator.

In this paper we present an approach to the design of overparameterized pole assignment compensators for SISO discrete-time systems which maintain stability in the presence of norm bounded structured parametric uncertainty in the plant model. The work is in a similar spirit to that in [7], but our use of norms is different and we design for nominal pole assignment. This gives a design which uses a convex optimization and allows some closed form results to be obtained.

II. PROBLEM FORMULATION

The z transform, \hat{h} , of a sequence, $\{h_i\}_{i=0}^\infty$, is defined by $\hat{h} = \sum_{i=0}^\infty h_i z^{-i}$. With this definition, a stable transfer function has all of its poles inside the unit disc of the complex plane. The symbol z^{-1} denotes the unit delay and polynomial, $X(z)$, of order n , is given by $X(z) = \sum_{i=0}^n x_i z^{-i}$.

In this work we consider the vector of coefficients of a polynomial as an element in an l_p normed space. We denote the vector of coefficients of $X(z)$ as x , with the l_p norm of a vector $x = \{x_i\}_{i=0}^\infty$ denoted as $\|x\|_p$, defined as

$$\|x\|_p = \left| \sum |x_i|^p \right|^{1/p}$$

for $1 \leq p < \infty$ with

$$\|x\|_\infty = \sup |x_i|$$

Consider a SISO discrete-time plant given by

$$\hat{y}_p = \frac{B(z)}{A(z)} \hat{u}_p$$

where \hat{y}_p is the plant output, u_p is its input, and $B(z)$ and $A(z)$ are coprime polynomials with $b_0 = 0$ and $a_0 = 1$. The plant has n_a finite poles $t_1, t_2, \dots, t_{n_a} \neq 0$. The plant is in a feedback control system with \hat{u}_p given by

$$\hat{u}_p = \frac{G(z)}{(1 - z^{-1})F(z)} w - \hat{y}_p$$

where w is the reference input and $G(z)$ and $F(z)$ are compensator polynomials. Note that for convenience a one degree of freedom (1-DOF) compensator is considered here, however the robustness results obtained apply equally for 2-DOF systems.

The system closed-loop poles are the roots of $V(z) = 0$, where

$$A(z)(1 - z^{-1})F(z) + B(z)G(z) = V(z). \quad (1)$$

For the usual pole-assignment controller design, the designer specifies $V(z)$ with all its roots stable and then solves (1) for the compensator polynomials, $F(z)$ and $G(z)$. The $F(z)$ and $G(z)$ which satisfy (1) are not unique but are parameterized as follows. If $F^0(z)$ and $G^0(z)$

Manuscript received December 21, 1993; revised July 1, 1994.

M. E. Halpern and R. J. Evans are with the Department of Electrical and Electronic Engineering, University of Melbourne and Cooperative Research Centre for Sensor Signal and Information Processing, Grattan St, Parkville, Victoria 3052, Australia.

R. D. Hill is with the Department of Mathematics, Royal Melbourne Institute of Technology, GPO Box 2476V, Melbourne, Victoria 3001, Australia.

IEEE Log Number 9408780

are a particular solution to (1), then all $F(z)$ and $G(z)$ which give the same $V(z)$ are given by

$$F(z) = F^0(-) + B(-) V(-) \quad (2)$$

and

$$G(-) = G^0(-) - (1 - z^{-1}) A(-) V(-) \quad (3)$$

where $V(-)$ is an arbitrary polynomial

Usually, $F^0(-)$ and $G^0(-)$ are chosen to be the minimum order unique polynomials satisfying (1), and $V(-)$ is incorporated to give some sort of performance improvement, through overparameterization of the compensator. Such a procedure gives proper controllers.

In the work presented here, $V(-)$ is chosen to reduce the effects of parametric plant perturbations on $V(-)$. We consider the plant as a nominal plant $\frac{B^0(-)}{A^0(-)}$ modified by perturbation polynomials $\Delta B(-)$ with coefficient vector $\Delta b = (0 \ \Delta b_1 \ \dots \ \Delta b_{n_{\Delta b}})$ and $\Delta A(-)$ with coefficient vector $\Delta a = (0 \ \Delta a_1 \ \dots \ \Delta a_{n_{\Delta a}})$ so that the perturbed plant is

$$\frac{B(-)}{A(-)} = \frac{B^0(-) + \Delta B(-)}{A^0(-) + \Delta A(-)}$$

The orders of $\Delta B(-)$ and $\Delta A(-)$ may be larger than those of $B^0(z)$ which has order n_{B^0} and $A^0(-)$ with order n_{A^0} . The order of the perturbed plant polynomials is

$$n_l = \max\{n_{b_l}, n_{\Delta b_l}\}$$

and

$$n_r = \max\{n_{r_0}, n_{\Delta r_l}\}$$

If the perturbed plant is used in conjunction with a compensator designed to give characteristic polynomial $V^0(-)$ with the nominal plant $B^0(-)/A^0(-)$, the closed-loop characteristic equation will be

$$V^0(-) + \Delta V(-) = 0$$

where from (1)

$$\Delta V(-) = G(-)\Delta B(-) + (1 - z^{-1})F(-)\Delta A(-) \quad (4)$$

Our approach involves designing the compensator to reduce the effect of $\Delta B(-)$ and $\Delta A(-)$ on the size of $\Delta V(-)$ thereby allowing larger plant perturbations while preserving stability.

III. ROBUSTNESS AGAINST A CLASS OF PLANT PERTURBATIONS

We consider the coefficient vectors of $\Delta B(-)$ and $\Delta A(-)$ as fixed but unknown elements in an l_p normed space.

Initially, we examine the case with $\Delta A(-) = 0$. In this case (4) reduces to

$$\Delta V(-) = G(-)\Delta B(-)$$

In terms of coefficients we then have

$$\Delta v_i = \sum_{j=0}^l q_j \Delta b_{i-j}, \quad i = 1, 2, \dots, n_r + n_{\Delta b_l}$$

so that Hölder's Inequality gives

$$\|\Delta v\|_\infty \leq \|q\|_l \|\Delta b\|_l \quad (5)$$

for all $p, q \in [1, \infty)$ satisfying $1/p + 1/q = 1$

We define γ to be the largest number such that

$$\|\Delta v\|_\infty < \gamma \Rightarrow V^0(-) + \Delta V(-) \text{ is stable}$$

A sufficient condition for stability of the perturbed system is $\|\Delta v\|_\infty < \gamma$. A sufficient condition for $\|\Delta v\|_\infty < \gamma$ is that for some $p, q \in [1, \infty)$ satisfying $1/p + 1/q = 1$ we have $\|q\|_l \|\Delta b\|_l < \gamma$.

This yields a key result which is a sufficiency test for stability if for some $p, q \in [1, \infty)$ satisfying $1/p + 1/q = 1$ we have

$$\|\Delta b\|_l < \gamma \quad (6)$$

then the perturbed system remains stable.

If "degree dropping," where the perturbations cause certain of the b_i to become zero, occurs, this does not invalidate (6).

We may interpret γ geometrically as half the length of side of the largest hypercube centered at v^0 , the vector of coefficients of the nominal closed-loop characteristic polynomial, and containing only stable polynomials. The value of γ depends only on v^0 and on n , the order of the perturbed closed-loop characteristic polynomial given by

$$n = \max\{n_b + n_r, n_r + n_l + 1\}$$

Vicino [10] has shown how to calculate γ . For the particular case where the nominal system is deadbeat with order n we have from Soh [11] that $\gamma = 1/n$.

Thus from (6) we can design a controller with robust stability against l_l perturbations to the plant numerator by minimizing $\|q\|_l$.

This minimization may be performed in several ways. One way is by calculating $G(-) = G^0(-) - (1 - z^{-1}) V^0(-) V(-)$ in terms of the unknown coefficients of $V(-)$ and then minimizing $\|q\|_l$ with respect to these coefficients. If $p = 1$ or $p = \infty$ the minimization is a linear program. If $p = 2$ there is a closed form solution.

Alternatively the minimization of $\|q\|_l$ may be formulated as the minimization of a norm subject to a set of linear interpolation constraints as follows

$$(7)$$

subject to

$$\sum_{i=1}^n q_i t_i = \frac{V^0(t_i)}{B^0(t_i)}, \quad i = 1, 2, \dots, n \quad (8)$$

and to

$$\sum_{i=0}^n q_i = \frac{V^0(1)}{B^0(1)} \quad (9)$$

where (9) is for an integrator pole. Here the t_i 's are the plant poles which are all known. They may be stable or unstable.

Writing the minimization in this form (7)–(9) yields the insight that "adding" more poles to the plant restricts the feasible set for the minimization and thus increases the value of $(\|q\|_l)_{\min}$, thereby tending to decrease the robustness to uncertainty in the plant numerator. In this respect the integrator imposed for performance, has the same kind of effect as an extra plant pole.

As the order of the system is increased through overparameterization of the compensator the values of both γ and of $(\|q\|_l)_{\min}$ become smaller. It is thus necessary to check that their quotient increases with overparameterization.

Another benefit of expressing the minimization as in (7)–(9) is that it enables the construction of a dual maximization problem, which we may use in some special cases to obtain an analytical solution to the minimization (7)–(9).

The maximization, dual to the primal minimization of (7)–(9), is given by

$$(\|q\|_l)_{\min} = \max_{\alpha, \beta} \left| \sum_{i=1}^n \alpha_i \frac{V^0(t_i)}{B^0(t_i)} + \alpha \frac{V^0(1)}{B^0(1)} \right|$$

subject to

$$\|q^*\|_l \leq 1$$

where

$$g_i^* = \sum_{j=1}^{n_g} \alpha_j t_j^{-i} + \alpha_{n_g+1}; \quad i = 0, \dots, n_g.$$

A. Closed Form Solutions

In this section we derive analytical expressions for $(\|\Delta b\|_q)_{\max}$ for a nominal deadbeat closed-loop system where the plant has only one pole $t_1 \in (-\infty, \infty)$ with $t_1 \neq 0$ and where there is no integrator in the loop.

For the problem of minimizing the l_p norm of a vector $x \in l_p$ subject to one linear constraint $\sum_{i=0}^{\infty} e_i x_i = m$ (which we write as $e \cdot x = m$), Hölder's Inequality yields

$$\begin{aligned} \min \|x\|_p &= \\ \text{s.t. } e \cdot x &= m \end{aligned} \quad (10)$$

where $p, q \in [1, \infty)$ and $1/p + 1/q = 1$.

Applying this result to the minimization problem of (7) gives

$$(\|g\|_p)_{\min} = \frac{1}{|B^0(t_1)| \|t_1^{(n_g)}\|_q}$$

where

$$= (1, t_1^{-1}, t_1^{-2}, \dots, t_1^{-n_g})$$

Since the closed-loop system is deadbeat, we have

$$\gamma = \frac{1}{n} = \frac{1}{n_b + n_g}$$

so that

$$(\|\Delta b\|_q)_{\max} = \frac{|B^0(t_1)| \|t_1^{(n_g)}\|_q}{n_b + n_g} \quad (11)$$

With $q = 1$, (11) gives

$$(\|\Delta b\|_1)_{\max} = \frac{|B^0(t_1)|}{(n_b + n_g)} \frac{(1 - |t_1|^{-(n_g+1)})}{(1 - |t_1|^{-1})}. \quad (12)$$

With $q = 2$, (11) gives

$$(\|\Delta b\|_2)_{\max} = \frac{|B^0(t_1)|}{n_b + n_g} |1 - t_1^{-(2n_g+2)}|^{1/2} \quad (13)$$

With $q = \infty$, (11) gives

if $|t_1| > 1$,

$$(\|\Delta b\|_{\infty})_{\max} = \frac{|B^0(t_1)|}{n_b + n_g} \quad (14)$$

and if $|t_1| \leq 1$,

$$(\|\Delta b\|_{\infty})_{\max} = \frac{|t_1^{-n_g} B^0(t_1)|}{n_b + n_g} \quad (15)$$

Here with no overparameterization, $n_g = 0$.

Equation (11) indicates that if the plant is stable, that is if $(|t_1| < 1)$, then the nominal deadbeat system can be made closed-loop stable against arbitrarily large perturbations of arbitrarily high order in the plant numerator by overparameterizing. We can also see this by inspecting (8), noting that the coefficients on the left-hand side of the constraint (8) may be made arbitrarily large by overparameterizing. This clearly holds for a plant with any number of only stable poles.

These closed form solutions (11)–(15) for $(\|\Delta b\|)_{\max}$ are each proportional to $B^0(t_1)$. A value of $B^0(t_1) = 0$ indicates a plant pole-zero cancellation. An unstable root common to $B^0(z)$ and $A^0(z)$ clearly means that the nominal plant is not stabilizable but a stable common root has no such effect. The closed form solutions are obtained on the basis of deadbeat pole placement, and this is not possible if there is a plant pole-zero cancellation.

B. Selection of Norms

Performing the design requires choosing the l_p norm of the compensator coefficient vector to be minimized. This should be done based on knowledge of the distribution of perturbation coefficient magnitudes to allow larger perturbations without losing a guarantee of stability. We do this by incorporating a scaling factor c into the perturbation so that

$$\Delta B(z) = c\delta(z).$$

This does not uniquely define c but if we keep $\delta(z)$ constant, we may usefully try to make c large. We then define $c_{(q)}$ according to

$$\|\Delta b\|_q = (\|\Delta b\|_q)_{\max} \Rightarrow c = c_{(q)}$$

and then attempt to find the value of q which maximizes $c_{(q)}$.

We consider only l_1, l_2 , and l_{∞} norms because these are the easiest to calculate. The results extend for the more general case of l_q norms with $q \in [1, \infty)$.

An important result behind the selection is that

$$(\|g\|_{\infty})_{\min} \leq (\|g\|_2)_{\min} \leq (\|g\|_1)_{\min}.$$

From (6) then we have that

$$(\|\Delta b\|_1)_{\max} \geq (\|\Delta b\|_2)_{\max} \geq (\|\Delta b\|_{\infty})_{\max}. \quad (16)$$

We illustrate the norm selection process by considering two extreme distributions of perturbation coefficient magnitudes. The first arises from a perturbation with only one nonzero coefficient, that is $\Delta B(z) = c z^{-k}$ where k is an unknown integer in the range $[1, n_{\Delta b}]$.

We wish to choose the l_p norm such that minimizing $\|g\|_p$ gives the largest value of $|c|$ with guaranteed stability. For this perturbation

$$\|\Delta b\|_1 = \|\Delta b\|_2 = \|\Delta b\|_{\infty} = |c|$$

so that

$$\|\Delta b\|_1 = (\|\Delta b\|_1)_{\max} \Rightarrow c_{(1)} = (\|\Delta b\|_1)_{\max},$$

$$\|\Delta b\|_2 = (\|\Delta b\|_2)_{\max} \Rightarrow c_{(2)} = (\|\Delta b\|_2)_{\max}$$

and

$$\|\Delta b\|_{\infty} = (\|\Delta b\|_{\infty})_{\max} \Rightarrow c_{(\infty)} = (\|\Delta b\|_{\infty})_{\max}.$$

Equation (16) implies that we obtain the largest value of $|c|$ by choosing $q = 1$ and $p = \infty$, that is we design by minimizing $\|g\|_{\infty}$.

At the other extreme, if the perturbation is such that $|\Delta b_i| = c$ for $i = 1, \dots, n_{\Delta b}$, then $\|\Delta b\|_1 = c n_{\Delta b}$, $\|\Delta b\|_2 = c \sqrt{n_{\Delta b}}$ and $\|\Delta b\|_{\infty} = c$ so that we have

$$\|\Delta b\|_1 = (\|\Delta b\|_1)_{\max} \Rightarrow c_{(1)} = \frac{(\|\Delta b\|_1)_{\max}}{n_{\Delta b}},$$

$$\|\Delta b\|_2 = (\|\Delta b\|_2)_{\max} \Rightarrow c_{(2)} = \frac{(\|\Delta b\|_2)_{\max}}{\sqrt{n_{\Delta b}}}$$

and

$$\|\Delta b\|_{\infty} = (\|\Delta b\|_{\infty})_{\max} \Rightarrow c_{(\infty)} = (\|\Delta b\|_{\infty})_{\max}.$$

For the simple case where the plant has only one pole $t_1 \in (-\infty, \infty)$ with $t_1 \neq 0$ and where the nominal closed-loop pole set is deadbeat, we can substitute for $(\|\Delta b\|_q)_{\max}$ from (7)–(9) to obtain

$$\begin{aligned} c_{(1)} &= \frac{|B^0(t_1)|}{(n_b + n_g)} \frac{(1 - |t_1|^{-(n_g+1)})}{(1 - |t_1|^{-1})} \frac{1}{n_{\Delta b}}, \\ c_{(2)} &= \frac{|B^0(t_1)|}{(n_b + n_g)} \frac{1 - t_1^{-(2n_g+2)}}{1 - t_1^{-2}} \frac{1}{\sqrt{n_{\Delta b}}} \end{aligned}$$

TABLE I
EFFECT OF OVERPARAMETERIZATION OF COMPENSATORS

Order of $G(z)$	$(\ L\ _1)_{\min}$	γ	$(\ \Delta b\ _1)_{\max}$
1	4.08654	1/3	0.0816
2	1.72476	0.25	0.1449
3	1.10385	0.2	0.1812
4	0.7871	1/6	0.2117

and

$$c_{(\infty)} = \frac{|B^0(t_1)|}{(n_b + n_q)}.$$

For $n_{\Delta b} \geq 2$, $n_g \geq 1$ and $|t_1| > 1$, we have $c_{(\infty)} > c_{(2)} > c_{(1)}$ and therefore can guarantee closed-loop stability for a larger value of γ if we minimize $\|g\|_1$.

C. Example

In this example we demonstrate how overparameterizing the compensator can increase the value of $(\|\Delta b\|_1)_{\max}$.

The nominal plant is

$$\frac{B^0(z)}{A^0(z)} = \frac{0.1333z^{-1} - 0.6667z^{-2}}{1 - 0.6667z^{-1}}$$

and so has a delay, a nonminimum phase zero and a stable pole. The nominal closed-loop pole set is to be deadbeat, and there is an integrator in the loop so (11) does not apply. The order $n_{\Delta b}$ of the perturbation is chosen to be $n_{\Delta b} = 2$ so that $n_b = \max(n_{b0}, n_{\Delta b}) = \max(2, 2) = 2$. Increasing the value of $n_{\Delta b}$ reduces the value of γ by increasing the value of n_b in the formula $\gamma = 1/(n_b + n_q)$ for a nominal deadbeat system. Table I shows the effect of overparameterization for this example. This table shows how the l_1 norm of the guaranteed safe perturbation on the plant numerator increases as the compensator is overparameterized.

D. Other Plant Perturbations

A similar approach can be taken for plant perturbations with $\Delta B(z) = 0$ and $\Delta A(z) \neq 0$.

For both $\Delta A(z)$, $\Delta B(z) \neq 0$ we may prefer on the basis of the distribution of perturbation coefficient magnitudes to consider Δb and Δa to be elements of different l_p normed spaces. We take $\Delta b \in l_{p_B}$ and $\Delta a \in l_{p_A}$ where $p_A, p_B \in [1, \infty)$. We then have, by a simple extension of (5) that

$$\|\Delta v\|_{\infty} \leq \|f\|_{p_A} \|\Delta a\|_{q_A} + \|g\|_{p_B} \|\Delta b\|_{q_B}$$

for all $p_A, q_A, p_B, q_B \in [1, \infty)$ satisfying $1/p_A + 1/q_A = 1$ and $1/p_B + 1/q_B = 1$.

Here, for convenience of notation, f is the coefficient vector of $(1 - z^{-1})F(z)$.

A sufficient condition for stability of the perturbed system is that

$$\|f\|_{p_A} \|\Delta a\|_{q_A} + \|g\|_{p_B} \|\Delta b\|_{q_B} < 1 \quad (17)$$

for some $p_A, q_A, p_B, q_B \in [1, \infty)$ satisfying $1/p_A + 1/q_A = 1$ and $1/p_B + 1/q_B = 1$.

The design approach then is to make $\|g\|_1$ and $\|f\|_1$ small in order to reduce the "gain" from each perturbation to Δv .

We cannot minimize both $\|g\|_1$ and $\|f\|_1$ independently with respect to $X(z)$ since the same value of $X(z)$ must be incorporated in both $G(z)$ and $F(z)$ as in (2) and (3). To design a compensator

to reduce the "gain" from each perturbation to Δv , we need to share the minimization effort between reducing $\|g\|_1$ and $\|f\|_1$.

There are a number of ways of doing this. One way is to formulate the compensator design as the minimization of a weighted sum of $\|f\|_1$ and $\|g\|_1$ with respect to $X(z)$ subject to the set of linear equality constraints obtained by equating like powers of z in

$$A^0(z)(1 - z^{-1})F^0(z) + B^0(z)G^0(z) = V^0(z). \quad (18)$$

Another way is to specify a numerical upper bound for either $\|g\|_1$ or $\|f\|_1$ (of course the value specified must be large enough to be achievable) and then to formulate the design as the minimization of an l_p norm of the other coefficient vector, subject to the constraints (18) and to that on the specified value of the norm of the first coefficient vector.

If only l_1 and l_{∞} norms are used in the previously mentioned schemes, the optimizations are linear programs.

By whatever means the minimization effort is apportioned, the perturbed system will be stable if (17) holds.

If either of the perturbations is of high order, it will reduce γ thereby reducing the maximum allowable value of the other perturbation.

IV. REMARK

In this paper we have calculated allowable bounds on the l_p norms of plant parameter perturbations such that $\|\Delta v\|_{\infty} < \gamma$ where γ has been chosen to guarantee stability. We could have selected a smaller value of γ sufficient to ensure that the perturbed closed-loop poles do not cross chosen regions surrounding the nominal closed loop poles. The approach could then be considered as design for robust pole assignment. Of course reducing the value of γ reduces allowable bounds on plant perturbations.

V. CONCLUSION

We have presented a framework which uses convex optimization to design overparameterized pole assignment compensators with guaranteed stability against l_p norm bounded plant parameter perturbations. The framework is flexible and a number of design problems may be formulated within it, depending on the nature of the parametric perturbations.

REFERENCES

- [1] M. A. Dahleh, "BIBO stability robustness in the presence of coprime factor perturbations," *IEEE Trans Automat Contr*, vol. 37, no. 3, pp. 352-355, 1992.
- [2] Y. C. Soh and R. J. Evans, "Characterization of allowable perturbations of a closed-loop system with pole-placement controller," *IEEE Trans Automat Contr*, vol. 33, no. 5, pp. 466-469, 1992.
- [3] V. L. Kharitonov, "Asymptotic stability of an equilibrium position of a family of systems of linear differential equations," *Differentsial Uravnen*, vol. 14, no. 11, pp. 2066-2088, 1978.
- [4] A. T. Fam and J. S. Meditch, "A canonical parameter space for linear systems design," *IEEE Trans Automat Contr*, vol. AC-23, no. 3, pp. 454-458, 1978.
- [5] C. B. Soh, C. S. Berger, and K. P. Dabke, "On the stability properties of polynomials with perturbed coefficients," *IEEE Trans Automat Contr*, vol. AC-30, no. 10, pp. 1033-1036, 1985.
- [6] R. M. Biernacki, H. Hwang, and S. P. Bhattacharya, "Robust stability with structured real parameter perturbations," *IEEE Trans Automat Contr*, vol. AC-32, no. 6, pp. 495-506, 1987.
- [7] L. H. Keel, S. P. Bhattacharya, and J. W. Howze, "Robust control with structured perturbations," *IEEE Trans Automat Contr*, vol. 33, no. 1, pp. 68-78, 1988.
- [8] C. S. Berger, "Robust controller design by minimization of the variation of the coefficients of the closed-loop characteristic equation," *IEE Proc Pt. D*, no. 3, 1984, pp. 103-107.
- [9] C. S. Berger, "Robust control of discrete systems," *IEE Proc Pt. D*, no. 4, 1989, pp. 165-170.

- [10] A. Vicino, "Some results on robust stability of discrete-time systems," *IEEE Trans. Automat. Contr.*, vol. 33, no. 9, pp. 844–847, 1988.
- [11] C. B. Soh, "Parameter space approach to control problems," Ph.D. dissertation, Monash University, 1986.

The Carathéodory–Fejér Problem and $\mathcal{H}_\infty/\ell_1$ Identification: A Time Domain Approach

Jie Chen and Carl N. Nett

Abstract—In this paper we study a worse-case, robust control oriented identification problem. This problem is in the framework of \mathcal{H}_∞ identification but the formulation here is more general. The available *a priori* information in our problem consists of a lower bound on the relative stability of the plant, an upper bound on a certain gain associated with the plant, and an upper bound on the noise level. The plant to be identified is assumed to lie in a certain subset in the space of \mathcal{H}_∞ , characterized by the assumed *a priori* information. The available experimental information consists of a corrupt finite output time series obtained in response to a known nonzero but otherwise arbitrary input. Our objective is to identify from the given *a priori* and experimental information an uncertain model which consists of a nominal model in \mathcal{H}_∞ and a bound on the modeling error measured in \mathcal{H}_∞ norm. We present both an identification algorithm and several explicit lower and upper bounds on the identification error. The proposed algorithm is in the class of interpolatory algorithms which are known to possess desirable optimality properties in reducing the identification error. This algorithm is obtained by solving an extended Carathéodory–Fejér problem via standard convex programming methods. Both the algorithm and error bounds can be applied to ℓ_1 identification problems as well.

I. INTRODUCTION

Recently, there has been a considerable amount of interest in system identification from a worst-case robust control oriented approach (see, e.g., [14], [34] and the references therein). A specific example is the \mathcal{H}_∞ identification problem initiated in [14]. In this problem, one typically needs to apply a sequence of complex sinusoidal inputs to estimate a sequence of point frequency samples of the plant transfer function, and one then seeks to identify, from the estimated frequency samples and certain assumed *a priori* information, an uncertain model. This problem has been studied extensively (cf. [14], [11], [12], [26], [27], [2], [21], [6]), and much of the research has been focused on identifying from a given sequence of corrupted point frequency samples a model consistent with \mathcal{H}_∞ control design.

Another problem that is closely related to that of the \mathcal{H}_∞ identification is more concerned with the task of identifying an uncertain model compatible with the ℓ_1 design methodology. This problem has been generally referred to as identification in ℓ_1 and has been studied in, e.g., [5], [16], [17], [20], [7], [32], [29]. In this problem, one typically assumes certain prior knowledge of the plant impulse response, and the identification procedure then proceeds with a sequence of time domain data.

Manuscript received October 3, 1994. This work was supported in part by ONR and NSF.

J. Chen is with the College of Engineering, University of California, Riverside, CA 92521-0425 USA.

C. N. Nett is with the United Technology Research Center, East Hartford, CT 06108 USA.

IEEE Log Number 9408784.

In the present paper, we consider a variant to the \mathcal{H}_∞ and ℓ_1 identification problems alluded to above. This problem is formulated in such a way that the assumed plant *a priori* information is identical to that in the framework of [14], while the assumed experimental information is more similar to those utilized in [5], [16], [20]. Our objective is to identify an uncertain model in \mathcal{H}_∞ which includes a nominal model in \mathcal{H}_∞ and a bound on the modeling error measured in \mathcal{H}_∞ norm. As such, this problem may be considered as a hybrid of the \mathcal{H}_∞ and ℓ_1 identification problems or merely a more general formulation of \mathcal{H}_∞ identification problems. We shall present both an interpolatory identification algorithm [31], [5], [6] and several explicit lower and upper bounds on identification error for this problem. Though our results are derived in the \mathcal{H}_∞ framework, they can be applied equally well to ℓ_1 identification problems.

The main difference between our approach and that initiated in [14] lies in that the identification here is conducted entirely in time domain. This has a number of implications. First, unlike in [14], there is no need in the present approach to transform the time response data to frequency samples. A more important merit with this approach is that the proposed interpolatory algorithm has a guaranteed optimality property for the overall identification problem. In contrast, the previously known interpolatory algorithms [6], [13] in the framework of [14] are suboptimal with respect to only the step of identification from point frequency samples; the overall optimality properties, however, are lost due to the transformation from the time response data to the frequency samples. Moreover, the main difference between this work and that of ℓ_1 identification [5], [16], [20], [32] lies in the assumed plant *a priori* information. Unlike that in ℓ_1 identification, the plant *a priori* information in this approach is assumed for the plant transfer function. Though it is possible to transform this *a priori* information to one on the plant impulse response (e.g., via Cauchy estimates) and hence to apply the ℓ_1 identification approach, the transformation will necessarily lead to undesirable conservatism and thus result in a loss of optimality properties. From these perspectives, it appears that the algorithm presented in this paper is the first algorithm with a guaranteed optimality property for the problem of \mathcal{H}_∞ identification. Finally, though our technique in deriving the algorithm is very similar to that of [28] (see also [33]), the two problems are clearly different in that the goal in this paper is to identify an uncertain model from certain given *a priori* information and experimental data, while the problem in [28] amounts to verifying the consistency of a given uncertain model to data.

A preliminary version of this paper has appeared previously as [4].

A. Organization

In Section II, we formulate the problem and provide a brief review on relevant notions in information-based complexity (IBC) [31]. In Section III, we present our identification algorithm. This algorithm relies on the solution of a related problem termed the consistency between the data and *a priori* information, which is concerned with the issue of validating the assumed *a priori* information and is reduced to a convex optimization problem. The construction of the algorithm together with the consistency problem amounts to solving an extended Carathéodory–Fejér [18] interpolation problem. Finally, explicit lower and upper bounds on the identification error are derived in Section IV, and the convergence properties of the proposed identification algorithm are established. The paper is concluded in Section V.

B. Notation

We shall use the following notations throughout this paper. Let \mathbf{Z} denote the set of integers, $\mathbf{Z}_+ := \{k \in \mathbf{Z} : k \geq 0\}$, and $\mathbf{Z}_{+,n} := \{k \in \mathbf{Z}_+ : k < n\}$. Let \mathbf{R} denote the set of real numbers, and \mathbf{R}^n the space of n dimensional real vectors. Moreover, let $\mathbf{R}_+ := \{x \in \mathbf{R} : x \geq 0\}$. The symbol $\bar{\sigma}(A)$ denotes the largest singular value of the matrix $A \in \mathbf{R}^{n \times n}$, and if A is a symmetric matrix, $\bar{\lambda}(A)$ denotes its largest eigenvalue. The transpose of $A \in \mathbf{R}^{n \times n}$ is denoted by A^T . For any positive $\rho \in \mathbf{R}_+$, let $\mathbf{D}_\rho := \{z \in \mathbf{C} : |z| < \rho\}$ and define the normed space

$$\mathcal{H}_{\infty,\rho} := \left\{ f: \mathbf{D}_\rho \rightarrow \mathbf{C} \mid f \text{ analytic in } \mathbf{D}_\rho \text{ and } \|f\|_{\infty,\rho} := \sup_{z \in \mathbf{D}_\rho} |f(z)| < \infty \right\}.$$

Furthermore, when $\rho = 1$, we shall replace \mathbf{D}_1 by \mathbf{D} , and $\mathcal{H}_{\infty,1}$ by \mathcal{H}_∞ , and so on. Note that the union of $\mathcal{H}_{\infty,\rho}$ for all $\rho > 1$ constitutes a linear space. We denote this linear space by $\mathcal{H}_+ := \bigcup_{\rho>1} \mathcal{H}_{\infty,\rho} \subset \mathcal{H}_\infty$. Recall also the normed spaces ℓ_∞ and ℓ_1 (see, e.g., [8]). We shall use the symbol $\|\cdot\|_\infty$ to denote the norm associated with each of the spaces ℓ_∞ and \mathcal{H}_∞ ; however, each of these uses should be clear from context. For each of the normed spaces $(X, \|\cdot\|_X)$ given above, we shall define $\bar{B}X(M) := \{x \in X : \|x\|_X \leq M\}$. Finally, for each $n \in \mathbf{Z}_+, n \geq 1$, we define a projection operator by $P_n: \ell_\infty \rightarrow \mathbf{C}^n$ such that $(P_n f)_k := f_k$ for all $k \in \mathbf{Z}_{+,n}$. Furthermore, we define the operator $T: \mathbf{C}^n \rightarrow \mathbf{C}^{n \times n}$ by

$$T(c_0, \dots, c_{n-1}) = \begin{bmatrix} c_0 & 0 & \cdots & 0 \\ & c_1 & c_0 & \cdots & 0 \\ & & c_{n-1} & c_{n-2} & \cdots & c_0 \end{bmatrix}$$

where $c := [c_0 \ \cdots \ c_{n-1}] \in \mathbf{R}^n$. Often, to simplify the notation, for a vector denoted by a lowercase letter, we shall use its capital version to denote the corresponding lower triangular Toeplitz matrix generated by T , i.e., $C := T(c_0, \dots, c_{n-1})$.

II. PROBLEM STATEMENT AND PRELIMINARY BACKGROUND

We consider the class of causal, linear, shift-invariant, exponentially stable, distributed parameter systems. This class of systems is identified with the linear space \mathcal{H}_+ . The plant to be identified is represented by an $h \in \mathcal{H}_+$ which is the standard z -Transform of the plant impulse response evaluated at $1/z$:

$$h(z) := \sum h_k z^k. \quad (2.1)$$

The plant *a priori* information consists of two positive constants $M > 0$ and $\rho > 1$, and the system to be identified is assumed to be an element of $\bar{B}\mathcal{H}_{\infty,\rho}(M)$. Note that the plant set $\bar{B}\mathcal{H}_{\infty,\rho}(M)$ is precisely the same as that considered in [14], [11], [12], [26], and [6], and interpretations on M and ρ may be found in [14].

The experimental procedure considered herein is similar to those in [16], [20], [5]. Specifically, the experimental data is obtained by applying an arbitrary nonzero input $u \in \ell_\infty$ to the plant $h \in \bar{B}\mathcal{H}_{\infty,\rho}(M)$, which generates an output $h * u$ as the convolution product of h and u . The output measurement is assumed to be corrupted by an additive bounded noise v , and the corrupted output is observed over a finite duration n . This experiment can be precisely described using the experiment operator

$$E_n(h, v) := P_n((h * u) + v)$$

where $v \in \bar{B}\ell_\infty(\epsilon)$, and $\epsilon > 0$ is a prescribed bound, and it represents the full *a priori* information on the output noise. Denote

$U := T(u_0, \dots, u_{n-1})$. Then, each data record $y \in \mathbf{R}^n$, where $y_k := (E_n(h, v))_k$, can be expressed as $y = U P_n h + P_n v$. The set of all possible data which may be generated from the experiment is

$$\mathcal{Y} := \{y \in \mathbf{R}^n : y = U P_n h + P_n v \text{ for some } (h, v) \in \bar{B}\mathcal{H}_{\infty,\rho}(M) \times \bar{B}\ell_\infty(\epsilon)\}.$$

For each $y \in \mathcal{Y}$, we define the set of indistinguishable systems (cf. [31], [22], [6]) by

$$\mathcal{P}(y) := \{h \in \bar{B}\mathcal{H}_{\infty,\rho}(M) : y = U P_n h + P_n v \text{ for some } v \in \bar{B}\ell_\infty(\epsilon)\}.$$

Given the above *a priori* and experimental information, our primary goal in this paper is to develop an uncertain model that is consistent with the \mathcal{H}_∞ design framework. This goal will be achieved by constructing an identification algorithm and by quantifying a corresponding identification error. An identification algorithm is a mapping¹ $A_n: \mathbf{R}^n \rightarrow \mathcal{H}_+$ which operates on the data record and produces an element $A_n(y) \in \mathcal{H}_+$. The local and global identification errors associated with an algorithm A_n are defined, as usual (cf. [31], [6]), by

$$e(A_n; \rho, M, \epsilon; y) := \sup_{h \in \mathcal{P}(y)} \|h - A_n(y)\|_\infty$$

and

$$c(A_n; \rho, M, \epsilon) := \sup_{y \in \mathcal{Y}} e(A_n; \rho, M, \epsilon; y)$$

respectively. Note that the global error provides a universal bound on the modeling error regardless of data records. Additionally, as in [14], [11], [12], [26], it will be used to define convergence property of identification algorithms. An Algorithm A_n is said to be convergent if it satisfies the property

$$\lim_{\substack{n \rightarrow \infty \\ \epsilon \rightarrow 0}} c(A_n; \rho, M, \epsilon) = 0.$$

We shall be particularly interested in a class of algorithms which are termed interpolatory algorithms (cf. [31], [23], [5], [6]). An algorithm A_n is said to be interpolatory if for any $y \in \mathcal{Y}$, $A_n(y) \in \mathcal{P}(y)$. Algorithms of interpolatory class are known to possess many desirable properties, among which is its optimality properties in reducing the identification error. To precisely state these properties, we introduce the concepts of the radius and diameter [31] for the set $\mathcal{P}(y)$, defined by

$$r(\mathcal{P}(y)) := \inf_{h \in \mathcal{H}_+} \sup_{p \in \mathcal{P}(y)} \|h - p\|_\infty$$

and

$$d(\mathcal{P}(y)) := \sup_{h, p \in \mathcal{P}(y)} \|h - p\|_\infty$$

respectively. More specifically, the quantities $r(\mathcal{P}(y))$ and $d(\mathcal{P}(y))$ are called local radius and diameter, and their global counterparts are defined by $r^* := \sup_{y \in \mathcal{Y}} r(\mathcal{P}(y))$ and $d^* := \sup_{y \in \mathcal{Y}} d(\mathcal{P}(y))$, respectively. It is well known that the radius $r(\mathcal{P}(y))$ coincides with the minimal possible, or optimal identification error [31], [22], [5], [6]. It is also well known that if $A_n(y) \in \mathcal{P}(y)$, then

$$r(\mathcal{P}(y)) \leq e(A_n; \rho, M, \epsilon; y) \leq d(\mathcal{P}(y)) \leq 2r(\mathcal{P}(y)). \quad (2.2)$$

¹In a broader sense, an identification algorithm acts on the input-output data pair (u, y) . To simplify the notation, we omit the dependence of all the relevant notions (e.g., $\mathcal{P}(y)$, $A_n(y)$, etc.) on the input and the *a priori* information.

In this sense, an interpolatory algorithm is said to be optimal to within a factor of two. Finally, let

$$s^* := \sup_{\substack{h \in \tilde{B}\mathcal{H}_{\infty, \rho}(M) \\ \|U^* P_n h\|_{\infty} \leq \epsilon}} \|h(z)\|_{\infty} = \sup_{h \in P(0)} \|h(z)\|_{\infty}. \quad (2.3)$$

Then, as in [6], [31], it can be shown that the global radius and diameter satisfy the inequalities

$$s^* \leq r^* \leq d^* = 2s^*. \quad (2.4)$$

Note that the above quantity s^* (defined with respect to \mathcal{H}_{∞} norm) constitutes also a lower bound for the global radius even when it is measured in l_1 norm.

III. AN IDENTIFICATION ALGORITHM

A. Consistency of Data and *a priori* Information

In the preceding definition of identification error, we have assumed implicitly that the set $P(y)$ is nonempty; otherwise, the identification error will become unbounded. Whether this assumption holds is concerned with the issue whether the given data record and the assumed *a priori* information are consistent. Precisely, the notion of the consistency between the data and *a priori* information is defined as follows.

Definition 3.1: Given the plant *a priori* information $(\rho, M) \in [1, \infty) \times [0, \infty)$ and the noise *a priori* information $\epsilon \in [0, \infty)$. A data record $y \in \mathbf{R}^n$ is said to be consistent with the plant and noise *a priori* information (ρ, M, ϵ) if $P(y) \neq \emptyset$.

In other words, the consistency between y and (ρ, M, ϵ) is equivalent to the existence of a model-noise pair $(h, v) \in B\mathcal{H}_{\infty, \rho}(M) \times B l_{\infty}(\epsilon)$ such that for this pair the data $y \in \mathbf{R}^n$ may be generated from the experiment $E_n(h, v)$. As such, the consistency problem amounts to establishing the existence of an analytic function $h \in B\mathcal{H}_{\infty, \rho}(M)$ whose first n Taylor coefficients satisfy the relation $y = U^* P_n h + P_n v$ for some $v \in B l_{\infty}(\epsilon)$. This characteristic bears a close resemblance to that of the Carathéodory-Fejér problem (cf. [9], [18]) and thus suggests the following necessary and sufficient condition.

Lemma 3.1: Let $(\rho, M) \in [1, \infty) \times [0, \infty)$ and $\epsilon \in [0, \infty)$. Then, $y \in \mathbf{R}^n$ is consistent with (ρ, M, ϵ) if and only if there exists a $v \in B l_{\infty}(\epsilon)$ such that

$$(Y^* - V^*)(Y^* - V^*)^T \leq M^2 U^* R^{-2} U^T \quad (3.1)$$

where $R := \text{diag}(1/\rho \cdots \rho^{n-1})$.

Proof: Let $f(z) := (1/M)h(z)$. Then, $f \in \tilde{B}\mathcal{H}_{\infty}(1)$ if and only if $h \in \tilde{B}\mathcal{H}_{\infty, \rho}(M)$. Note that for any $z \in \mathbf{D}$, $h(z) = \sum_{k=0}^{\infty} h_k \rho^k z^k$. Hence, $f(z)$ can be expanded as $f(z) = \sum_{k=0}^{\infty} f_k z^k$ with $f_k = (1/M)\rho^k h_k$ for all $k \in \mathbf{Z}_+$. It follows that $P_n f = (1/M)R P_n h$, and that $y = U^* P_n h + P_n v$ if and only if $y = M U^* R^{-1} P_n f + P_n v$. Therefore, $P(y) \neq \emptyset$ if and only if there exists a function $f \in \tilde{B}\mathcal{H}_{\infty}(1)$ such that $y = M U^* R^{-1} P_n f + P_n v$ for some $v \in B l_{\infty}(\epsilon)$. The result then follows from utilizing an extended Carathéodory-Fejér theorem given in [30, p. 26]. \square

Based upon Lemma 3.1, we give below two equivalent necessary and sufficient conditions that can be readily computed. To simplify our presentation, we shall assume, with no loss of generality, that $u_0 \neq 0$. Under this assumption, U is invertible and its inverse is also a lower triangular Toeplitz matrix [5].

Theorem 3.1: Let $(\rho, M) \in [1, \infty) \times [0, \infty)$ and $\epsilon \in [0, \infty)$. Assume also that $u_0 \neq 0$. Then, $y \in \mathbf{R}^n$ is consistent with (ρ, M, ϵ) if and only if

$$\min_{v \in B l_{\infty}(\epsilon)} \bar{\sigma}(R U^{-1}(Y^* - V^*)) \leq M. \quad (3.2)$$

Proof: Since U is invertible, condition (3.1) is equivalent to that

$$(R U^{-1}(Y^* - V^*))(R U^{-1}(Y^* - V^*))^T \leq M^2 I. \quad (3.3)$$

This, however, is equivalent to that $\bar{\sigma}(R U^{-1}(Y^* - V^*)) \leq M$. Hence, the proof is completed. \square

Similar to the results given in [6], [28], Theorem 3.1 shows that the consistency problem can be solved as a convex program. This is clear by noting the fact that the largest singular value $\bar{\sigma}(A)$ is a convex function of A , and that the matrix $R U^{-1}(Y^* - V^*)$ is an affine function of $v_k, k \in \mathbf{Z}_{n+}$. As in [6], it is easy to derive necessary or sufficient conditions from this result.

Theorem 3.2: Let $(\rho, M) \in [1, \infty) \times [0, \infty)$ and $\epsilon \in [0, \infty)$. Furthermore, denote

$$A(v) := \begin{bmatrix} -M^2 U^* R^{-2} U^T & -(Y^* - V^*) \\ -(Y^* - V^*)^T & -I \end{bmatrix}.$$

Then, $y \in \mathbf{R}^n$ is consistent with (ρ, M, ϵ) if and only if

$$\min_{v \in B l_{\infty}(\epsilon)} \lambda(A(v)) \leq 0. \quad (3.4)$$

Proof: By Lemma 3.1 and a well-known fact concerning the Schur complement (see, e.g., [15]), we assert that (3.1) is equivalent to the condition

$$\begin{bmatrix} M^2 U^* R^{-2} U^T & (Y^* - V^*) \\ (Y^* - V^*)^T & I \end{bmatrix} > 0.$$

The latter condition, however, is equivalent to that $\lambda(A(v)) \leq 0$. Hence, the proof is completed. \square

Minimization problem (3.4) can also be solved as a convex program. This follows from the fact that the largest eigenvalue $\lambda(A)$ of a symmetric matrix A is a convex function of A (see, e.g., [15]), and that the matrix $A(v)$ is an affine function of $v_k, k \in \mathbf{Z}_{n+}$. Each of problems (3.2) and (3.4) belongs to the class of constrained nondifferentiable convex programs. Numerically, problem (3.4) is more amenable to implementation, as it possesses the standard form for this class of problems, for which the techniques developed in e.g., [3], [25] can be directly applied. Note also that in this result, the assumption $u_0 \neq 0$ is dropped.

Finally, as in [5], [6], we comment on the problems of computing the minimal *a priori* information levels that are required for the *a priori* information and data record to be consistent. The utility of these minimal information levels lies in that they help build up one's confidence in the assumed *a priori* information, particularly when the computed levels are significantly lower than those assumed. Toward this end, one problem is compute, given ρ and ϵ

$$M^* := \min\{M: P(y) \neq \emptyset\}. \quad (3.5)$$

Clearly, as a by-product of the data consistency problem, M^* is solved as

$$M^* = \min_{v \in B l_{\infty}(\epsilon)} \bar{\sigma}(R U^{-1}(Y^* - V^*)).$$

Another problem is to compute, given M and ρ

$$\epsilon^* = \min\{\epsilon: P(y) \neq \emptyset\}. \quad (3.6)$$

This quantity may be computed by solving a convex program analogous to (3.4)

$$\epsilon^* = \min\{\epsilon: A(v) \leq 0, \quad |v_k| \leq \epsilon, \quad k \in \mathbf{Z}_{n+}\}.$$

Alternatively, it can be solved by computing $\epsilon^* = \min\{\epsilon: f(\epsilon) \leq M\}$; where $f(\cdot): \mathbf{R}_+ \rightarrow \mathbf{R}_+$ is defined by

$$f(\epsilon) := \min_{v \in B l_{\infty}(\epsilon)} \bar{\sigma}(R U^{-1}(Y^* - V^*)).$$

As in [6], it can be readily shown that $f(\cdot)$ is continuous and nonincreasing. Hence, ϵ^* may be computed easily by using, e.g., the bisection method, given the fact that $f(\epsilon)$ may be solved for each $\epsilon \in \mathbf{R}_+$. Finally, it is also of interest to compute, given M and ϵ

$$\rho^* := \max\{\rho: \mathcal{P}(y) \neq \emptyset\} \quad (3.7)$$

which can be solved by computing $\rho^* = \max\{\rho: g(\rho) \leq M\}$, where $g(\cdot): [1, \infty) \rightarrow \mathbf{R}_+$ is defined by

$$g(\rho) := \max_{v \in H(\infty, \epsilon)} \bar{\sigma}(R U^{-1}(Y - V)).$$

Clearly, $g(\cdot)$ is continuous. It is also easy to show that $g(\cdot)$ is nondecreasing. Indeed, let $\rho_1 \leq \rho_2$, and denote $R_1 := \text{diag}(1/\rho_1 \cdots 1/\rho_1^{n-1})$ and $R_2 := \text{diag}(1/\rho_2 \cdots 1/\rho_2^{n-1})$. Then, for any $v \in \bar{B}(\infty, \epsilon)$

$$\begin{aligned} \bar{\sigma}(R_1 U^{-1}(Y - V)) &= \bar{\sigma}(R_1 R_2^{-1} R_2 U^{-1}(Y - V)) \\ &\leq \bar{\sigma}(R_1 R_2^{-1}) \bar{\sigma}(R_2 U^{-1}(Y - V)) \\ &= \bar{\sigma}(R_2 U^{-1}(Y - V)). \end{aligned}$$

This implies that $g(\rho_1) \leq g(\rho_2)$. Since $g(\cdot)$ is continuous and nondecreasing, problem (3.7) may be solved by using the bisection method as well.

B. An Interpolatory Algorithm

As an immediate consequence, the solution of the data consistency problem yields as a by-product an interpolatory algorithm. Indeed, once the consistency problem is solved, we may construct immediately an "interpolating" function $h \in B\mathcal{H}_{\infty, \rho}(M)$ using the Schur algorithm [9] or a procedure of [1] (see also [18]), such that its first n Taylor coefficients satisfy the relation $y = U P_n h + P_n v^*$ with v^* being the solution of the consistency problem. Clearly, such an interpolating function lies in the set $\mathcal{P}(y)$, and it can be constructed using the following interpolatory algorithm.

Algorithm 3.1

- **Step 1:** Solve minimization problem (3.4). Let the minimizer be denoted by

and the minimum be denoted by λ^* . Accordingly, write $V^* := T(v_0^*, \dots, v_{n-1}^*)$. If $\lambda^* > 0$, stop. The data and *a priori* information are not consistent. Otherwise go to Step 2.

- **Step 2:** Compute the matrix $H^* := (1/M) R U^{-1}(Y - V^*)$. If $\lambda^* < 0$, go to Step 3a. Otherwise compute the rank of $I - H^* H^{*T}$ and go to Step 3b.
- **Step 3a:** Compute the polynomials

$$\begin{aligned} p(z) &:= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} T(1, \dots, z^{n-1}) (I - H^* H^{*T})^{-1} \\ &\quad \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &\quad \vdots \\ &\quad 0 \\ &\quad 1 \end{aligned} \quad (3.8) \end{aligned}$$

$$\begin{aligned} q(z) &:= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} T(p_{n-1}(z), \dots, p_0(z)) \\ &\quad \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &\quad \times (I - H^* H^{*T})^{-1} \end{aligned} \quad (3.9)$$

where $p_k(z) := \sum_{i=0}^k h_i z^{k-i}$ for all $k \in \mathbf{Z}_{+,n}$, and $P_n h^* := U^{-1}(y - P_n v^*)$. The proposed algorithm A_n is given by

$$A_n(y)(z) := M \frac{(\frac{z}{\rho})^n p(\frac{z}{\rho})c + (\frac{z}{\rho})^{n-1} q(\frac{z}{\rho})}{(\frac{z}{\rho})^n q(\frac{z}{\rho})c + p(\frac{z}{\rho})}$$

where $|c| \leq 1$ is an arbitrary real constant.

Step 3b: Let $\text{rank}(I - H^* H^{*T}) = m$. Also, let I be the $(m+1) \times (m+1)$ identity matrix and

$$H := [I \ 0] H^* \begin{bmatrix} I \\ 0 \end{bmatrix}.$$

Solve a vector $c^T = [c_0 \cdots c_m]$ from the equation $(I - H H^T)c = 0$. The proposed algorithm A_n is given by

$$A_n(y)(z) := M \frac{c_0 \rho^m + c_1 \rho^{m-1} z + \cdots + c_m z^m}{c_m \rho^m + c_{m-1} \rho^{m-1} z + \cdots + c_0 z^m}.$$

The above algorithm, as those developed in [6], consists of two parts which solve a consistency problem first and a standard Carathéodory–Fejér problem subsequently. The key to the algorithm is clearly Step 1, which determines whether the measured data and the assumed *a priori* information are consistent. When $\lambda^* > 0$, the consistency problem admits no solution and the algorithm stops. When $\lambda^* \leq 0$, the data and *a priori* information are consistent and the algorithm proceeds to find an interpolating function $f \in \bar{B}\mathcal{H}_{\infty}(1)$. This function is found in either Step 3a or Step 3b and is then transformed to an interpolating function $h \in \bar{B}\mathcal{H}_{\infty, \rho}(M)$ via the transformation $h(z) := M f(z/\rho)$. When $\lambda^* = 0$, the interpolating function is unique [18] and is found in Step 3b. The algorithm produces a real rational transfer function with an order of m . When $\lambda^* < 0$, the interpolating function is nonunique [18] and can be parameterized by an analytic function in $B\mathcal{H}_{\infty}(1)$. Step 3a gives one specific choice of such functions by selecting the free parameter as a real constant c , which yields a real rational transfer function with an order of n . The construction in Step 3a and Step 3b follows from a modified procedure based on [1]. Note that the algorithm may be numerically ill-conditioned when $I - H^* H^{*T}$ is nearly singular. This, however, should not pose a serious problem for a well-conditioned algorithm may be obtained by resorting to an alternative procedure given in [1].

IV. BOUNDS ON IDENTIFICATION ERROR

In this section we provide several explicit lower and upper bounds for the identification error associated with the interpolatory algorithm proposed above. These bounds are derived for the global radius and diameter of the set $\mathcal{P}(y)$; therefore, they are applicable to any interpolatory algorithms. Our first result is a lower bound upon the global radius of $\mathcal{P}(y)$. As noted at the close of Section II, this result also serves as a lower bound for the corresponding ℓ_1 identification problem.

Theorem 4.1 Let $\|P_n u\|_{\infty} \neq 0$. Then

$$r^* \geq \min \left\{ M, \frac{\epsilon}{\|P_n u\|_{\infty}} \right\}. \quad (4.1)$$

Furthermore, if $\epsilon/\|P_n u\|_{\infty} < M$, then

$$> \frac{(M/\rho^n) + (\epsilon/\|P_n u\|_{\infty})}{1 + (1/M\rho^n)(\epsilon/\|P_n u\|_{\infty})}. \quad (4.2)$$

Proof: To establish Part i), consider the function

$$h(z) := \min \left\{ M, \frac{\epsilon}{\|P_n u\|_{\infty}} \right\}.$$

Clearly, $h \in B\mathcal{H}_{\infty, \rho}(M)$. Note also that $\|U P_n h\|_{\infty} \leq \max_{k \in \mathbf{Z}_{+,n}} (|u_k|/\|P_n u\|_{\infty}) \epsilon = \epsilon$. Hence $h \in \mathcal{P}(0)$. The result then follows by using (2.4). Suppose now that $(\epsilon/\|P_n u\|_{\infty}) < M$ and consider the function

$$h(z) := M \frac{(z^n/\rho^n) + (\epsilon/\|P_n u\|_{\infty})(1/M)}{1 + (z^n/\rho^n)(\epsilon/\|P_n u\|_{\infty})(1/M)}.$$

Since $(\epsilon/\|P_n u\|_{\infty}) < M$, h is analytic in \mathbf{D}_{ρ} . Furthermore, we have $|h(z)| \leq M$ for any $z \in \mathbf{D}_{\rho}$. The latter assertion follows from the

well-known fact that for any $a, b \in \mathbf{D}$ the inequality (see, e.g., [10, p. 4])

$$\frac{a-b}{1-ba} \leq \frac{|a|+|b|}{1+|a| \cdot |b|} \leq 1. \quad (4.3)$$

holds. As a result, $h \in \bar{B}\mathcal{H}_{\infty, \rho}(M)$. Note now that h can be expanded in Taylor series as

$$h(z) = \frac{\epsilon}{\|P_n u\|_{\infty}} + M \left(1 - \left(\frac{\epsilon}{M\|P_n u\|_{\infty}} \right)^2 \right) \frac{z^n}{\rho^n} - \dots$$

It follows that $\|U P_n h\|_{\infty} = \max_{k \in \mathbf{Z}_{+,n}} (|u_k|/\|P_n u\|_{\infty})\epsilon = \epsilon$. Hence again, $h \in \mathcal{P}(0)$. To complete the proof, we claim that

$$= \frac{(M/\rho^n) + (\epsilon/\|P_n u\|_{\infty})}{1 + (1/\rho^n)(\epsilon/\|P_n u\|_{\infty})(1/M)}.$$

Indeed, according to (4.3), we have

$$\|h(z)\|_{\infty} \leq \frac{(M/\rho^n) + (\epsilon/\|P_n u\|_{\infty})}{1 + (1/\rho^n)(\epsilon/\|P_n u\|_{\infty})(1/M)}.$$

The upper bound, however, is attained at $z = 1$. The proof for Part ii) is now completed by using (2.4). \square

The significance of the lower bound (4.1) lies in that it provides an irreducible level of identification error in terms of the quantity $\epsilon/\|P_n u\|_{\infty}$, which, as in [14], may be interpreted as a noise-to-signal ratio. The lower bound (4.2) provides an estimate tighter than (4.1) when $(\epsilon/\|P_n u\|_{\infty}) < M$. Note that in the noise free case ($\epsilon = 0$), this lower bound reduces to $r^* \geq M/\rho^n$, which together with a result of [24] yields

$$\frac{M}{\rho^n} \leq r^* \leq d^* \leq \frac{2M}{\rho^n}.$$

This inequality and (4.2) suggest that the global identification error will be inherently large when M is large and ρ is close to one.

We now present an upper bound for the global diameter of the set $\mathcal{P}(y)$. As in [14], [6], this bound is also useful for establishing the convergence property of Algorithm 3.1 in \mathcal{H}_{∞} . Furthermore, it shows that the algorithm is convergent in ℓ_1 , and when combined with the error bound, it leads to an uncertain model consistent with ℓ_1 design as well. We shall assume that $u_0 \neq 0$. Hence, we may write $W := T(u_0, \dots, w_{n-1}) = U^{-1}$. Note that the vector $w := [w_0 \dots w_{n-1}]^T$ can be calculated explicitly from U^{-1} .

Theorem 4.2: Assume that $u_0 \neq 0$. Let $l \in \mathbf{Z}_+$ be the largest integer k such that

$$\sum_{i=0}^l |w_i| \leq \frac{M}{\rho^l}. \quad (4.4)$$

Denote further that $l^* := \min\{l, n\}$. Then

$$d^* \leq 2 \left(\epsilon \sum_{k=0}^{l^*} (l^* + 1 - k) |w_k| + \frac{M}{\rho^{l^*-1}(\rho - 1)} \right). \quad (4.5)$$

In addition, any interpolatory algorithm is convergent.

Proof: For any $h \in \mathcal{P}(0)$, it can be expressed as $h = \sum_{k=0}^{\infty} h_k z^k$, where h_k satisfies the condition $U P_n h = P_n v$ for some $v \in \bar{B}\ell_{\infty}(\epsilon)$. It follows that $h_k = \sum_{i=0}^{\infty} w_{k-i} v_i$ for all $k \in \mathbf{Z}_{+,n}$. Additionally, by Cauchy's estimates (see, e.g., [19], p. 137), we have $|h_k| \leq M/\rho^k$ for all $k \in \mathbf{Z}_+$, since $h \in \bar{B}\mathcal{H}_{\infty, \rho}(M)$. Therefore,

it follows that

$$\begin{aligned} & \sup_{n \in \mathcal{P}(0)} \|h(z)\|_1 \\ &= \max \left\{ \|h\|_1; P_n h = U^{-1} P_n v, \right. \\ & \quad \left. |h_k| \leq M/\rho^k, k \in \mathbf{Z}_+, v \in \bar{B}\ell_{\infty}(\epsilon) \right\} \\ &= \max \left\{ \|W^* P_n v\|_1, \sum_{k=0}^l w_{k-i} v_i \leq \frac{M}{\rho^k}, \right. \\ & \quad \left. k \in \mathbf{Z}_{+,n}, v \in \bar{B}\ell_{\infty}(\epsilon) \right\} + \sum_{k=n}^{\infty} \frac{M}{\rho^k} \\ &\leq \sum_{k=0}^{n-1} \min \left\{ \epsilon \sum_{i=0}^k |w_i|, \frac{M}{\rho^k} \right\} + \sum_{k=n}^{\infty} \frac{M}{\rho^k} \\ &= \epsilon \sum \sum |w_i| + \sum_{k=l^*+1}^{\infty} \frac{M}{\rho^k} \\ &= \epsilon \sum_{k=0}^{l^*} (l^* + 1 - k) |w_k| + \frac{M}{\rho^{l^*}(\rho - 1)}. \end{aligned}$$

By a well-known fact [8], however, $\|h\|_{\infty} \leq \|h\|_1$. Consequently, the upper bound (4.5) may be established by using (2.4). To verify the convergence property of interpolatory algorithms, we show that $d^* \rightarrow 0$ as $n \rightarrow \infty$ and $\epsilon \rightarrow 0$. Indeed, it follows from the above proof that

$$\begin{aligned} d^* &\leq 2 \left(\sum_{k=0}^{n-1} \min \left\{ \epsilon \sum_{i=0}^k |w_i|, \frac{M}{\rho^k} \right\} + \sum_{k=n}^{\infty} \frac{M}{\rho^k} \right) \\ &\leq 2 \left(l^* \frac{M}{\rho^{l^*}} + \frac{M}{\rho^{l^*}(\rho - 1)} \right) \end{aligned}$$

where the last upper bound converges to zero when $l^* \rightarrow \infty$. Since $l^* = \min\{l, n\}$, and since $\epsilon \sum_{i=0}^{l^*} |w_i| > M/\rho^{l^*+1}$, it follows that $l^* \rightarrow \infty$ as $n \rightarrow \infty$ and $\epsilon \rightarrow 0$. Therefore, $d^* \rightarrow 0$ as $n \rightarrow \infty$ and $\epsilon \rightarrow 0$. This completes our proof. \square

In the remainder of this section we present an additional upper bound for the global diameter d^* . This result may be combined with Theorem 4.2 to yield an improved estimate of the global identification error in the \mathcal{H}_{∞} (but not in ℓ_1) case. Note that if $u_0 \neq 0$ and $\epsilon/|u_0| \geq M$, then d^* can always be bounded by $d^* \leq 2M \leq 2(\epsilon/|u_0|)$. This follows by examining the quantity s^* defined by (2.3), and by noting the fact that $s^* \leq \sup_{h \in \mathcal{H}_{\infty, \rho}(M)} \|h\|_{\infty} = M$. Hence, it suffices to consider the case $\epsilon/|u_0| < M$. The following lemma, given in [6], is useful in our derivation.

Lemma 4.1: Let $f: [0, 1] \times [0, 1] \rightarrow \mathbf{R}_+$ be defined by

$$f(x, y) := \frac{x+y}{1+xy}.$$

Then, $f(x, y)$ is monotonically increasing with respect to both x and y .

Theorem 4.3: Assume that $u_0 \neq 0$ and $\epsilon/|u_0| < M$. Then

$$d^* \leq 2 \frac{(\epsilon/|u_0|) + M(\gamma/\rho)}{1 + (\epsilon/M|u_0|)(\gamma/\rho)} \quad (4.6)$$

where

$$\gamma := \sum_{k=0}^{n-2} \min \left\{ \alpha_k, \frac{1}{\rho^k} \right\} + \frac{1}{\rho^{n-2}(\rho - 1)}$$

and for $k \in \mathbf{Z}_{+,n-1}$, α_k can be computed iteratively as

$$\alpha_k := \epsilon \frac{\rho M \sum_{i=0}^{k+1} |w_i| + (\epsilon/|u_0|) \sum_{i=1}^k \alpha_{k-i} \sum_{j=1}^i |w_j|}{M^2 - (\epsilon/|u_0|)^2}.$$

Proof Consider any $h \in \mathcal{P}(0)$. Then, $h \in \bar{\mathcal{B}}\mathcal{H}_{\infty, \rho}(M)$ and it satisfies the condition $h_k = \sum_{i=0}^k u_{k-i} \rho^i$, $k \in \mathbb{Z}_{+, n}$ for some $\rho \in \bar{\mathcal{B}}\mathcal{H}_{\infty}(\epsilon)$. Let f be defined by $f(\cdot) = (1/M)h(\rho\cdot)$. It follows that $f \in \bar{\mathcal{B}}\mathcal{H}_{\infty}(1)$ and $f_k = (1/M)\rho^k h_k$ for $k \in \mathbb{Z}_{+, n}$. Since h interpolates h_k , $k \in \mathbb{Z}_{+, n}$ in $\bar{\mathcal{B}}\mathcal{H}_{\infty, \rho}(M)$, f interpolates f_k , $k \in \mathbb{Z}_{+, n}$ in $\bar{\mathcal{B}}\mathcal{H}_{\infty}(1)$. Furthermore, since $\epsilon/|u_0| < M$, we have $|f_0| = (1/M)|h_0| = (1/M)|u_0 \rho^0| = (1/M)|u_0|/\rho \leq \epsilon/(M|u_0|) < 1$. Therefore, by Schur's algorithm (see, e.g., [9]), f may be expressed as

$$f(\cdot) = \frac{f_0 + \rho g(\cdot)}{1 + f_0 \rho g(\cdot)}$$

where $g \in \bar{\mathcal{B}}\mathcal{H}_{\infty}(1)$ satisfies the condition

$$g_k = \frac{f_{k+1} + f_0 \sum_{i=1}^k g_{k-i}}{1 - |f_0|^2}, \quad k \in \mathbb{Z}_+$$

By inequality (4.3) it follows that for any $z \in \mathcal{D}$ $|f(\cdot)| \leq (|f_0| + |\rho g(\cdot)|)/(1 + |f_0 \rho g(\cdot)|)$. This implies that for any $\cdot \in \mathcal{D}_\rho$

$$|h(\cdot)| = M|f(\cdot/\rho)| \leq M \frac{|f_0| + |\rho g(\cdot/\rho)|}{1 + |f_0| |\rho g(\cdot/\rho)|}$$

Therefore, by Lemma 4.1, we have

$$\begin{aligned} \gamma^* &= \sup_{f \in \mathcal{P}(0)} \|h(\cdot)\|_{\infty} \\ &\leq M \frac{(\epsilon/|Mu_0|) + (1/\rho) \sup_{g \in \Omega} \|g(\cdot)\|_{\infty}}{1 + (\epsilon/|Mu_0|)(1/\rho) \sup_{g \in \Omega} \|g(\cdot)\|_{\infty}} \end{aligned}$$

where $g(\cdot) = g(\cdot/\rho) \in \bar{\mathcal{B}}\mathcal{H}_{\infty, \rho}(1)$ and $\Omega = \{g \in \bar{\mathcal{B}}\mathcal{H}_{\infty, \rho}(1) : g_k = q_k/\rho^k, k \in \mathbb{Z}_{+, n-1}, \rho \in \bar{\mathcal{B}}\mathcal{H}_{\infty}(\epsilon)\}$. As a result, for any $g \in \Omega$, we have

$$\begin{aligned} q_k &= (1/\rho^k) \frac{f_{k+1} + f_0 \sum_{i=1}^k g_{k-i}}{1 - |f_0|^2} \\ &= \frac{\rho M h_{k+1} + h_0 \sum_{i=1}^k q_{k-i}}{M^2 - |h_0|^2} \end{aligned}$$

It is clear that $|q_0| \leq \alpha_0$. Suppose that $|q_i| \leq \alpha_i$ for $i \in \mathbb{Z}_{+, n-1}$. Then, we have

$$\begin{aligned} |q_k| &\leq \epsilon \frac{\rho M \sum_{i=0}^{k+1} |u_i| + (\epsilon/|u_0|) \sum_{i=1}^k |q_{k-i}| \sum_{j=1}^k |u_j|}{M^2 - (\epsilon/|u_0|)^2} \\ &\leq \epsilon \frac{\rho M \sum_{i=0}^{k+1} |u_i| + (\epsilon/|u_0|) \sum_{i=1}^k \alpha_{k-i} \sum_{j=1}^k |u_j|}{M^2 - (\epsilon/|u_0|)^2} \\ &= \alpha_k \end{aligned}$$

Therefore, we conclude that $|q_k| \leq \alpha_k$ for all $k \in \mathbb{Z}_{+, n-1}$. It follows as in the proof of Theorem 4.2 that

$$\sup_{g \in \Omega} \|g(\cdot)\|_{\infty} \leq \sum_{k=0}^{n-2} \min \left\{ |q_k|, \frac{1}{\rho^k} \right\} + \frac{1}{\rho^{n-2}(\rho-1)} \leq \gamma$$

Hence again, by Lemma 4.1, we conclude that

$$\gamma^* \leq \frac{(\epsilon/|u_0|) + M(\gamma/\rho)}{1 + (\epsilon/|Mu_0|)(\gamma/\rho)}$$

The proof may now be completed by applying (2.4). \square

Note that the upper bound in (4.6) also indicates that any interpolatory algorithm is convergent. Indeed, a technique similar to the proof of Theorem 4.2 may be employed to show that γ approaches to zero when $n \rightarrow \infty$ and $\epsilon \rightarrow 0$. This upper bound is derived by combining the Schur's iterative algorithm [9] and the technique used in the proof of Theorem 4.2, which essentially involves approximating a function $h \in \mathcal{P}(0)$ via a first order rational function.

V CONCLUSION

We have presented an identification algorithm for a worst-case \mathcal{H}_{∞} identification problem. The main feature about this algorithm is that it is an interpolatory algorithm constructed based on time domain data. A direct consequence of this feature is that the algorithm is optimal to within a factor of two. This property is not shared by the previously available algorithms for \mathcal{H}_{∞} identification problems. Note also that the algorithm is interpolatory for the corresponding ℓ_1 identification problem as well.

A main disadvantage of this algorithm is that it requires solving a convex optimization problem with a size equal to the number of data. This may be computationally demanding. By trading off the optimality properties, it is possible to lessen this computational complexity. How to achieve a judicious trade-off between the optimality and complexity is a difficult issue and appears to be a worthy research topic. Future problems also exist in the aspects of, e.g., derivation of lower order models, identification for multivariable systems and issues related to input design and sample complexity [7], [29], [17].

REFERENCES

- [1] V. M. Adamjan, D. Z. Arov, and M. G. Krein, "Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur Takagi problem," *Math. USSR Sbornik*, vol. 15, no. 1, pp. 31-73, 1971.
- [2] F. Bai and S. Raman, "A linear robustly convergent interpolatory algorithm for system identification," in *Proc. 1992 Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 3165-3169.
- [3] S. Boyd and L. Li Ghaoui, "Method of centers for minimizing generalized eigenvalues," *Stanford Univ. Tech. Rep.*, Apr. 1992.
- [4] J. Chen and C. N. Nett, "The Carathéodory-Fejér problem and \mathcal{H}_{∞} identification: A time domain approach," in *Proc. 32nd IEEE Conf. Decis. Contr.*, San Antonio, TX, Dec. 1993, pp. 68-73.
- [5] J. Chen, C. N. Nett, and M. K. H. Fan, "Optimal nonparametric system identification from arbitrary corrupt finite time series: A control oriented approach," in *Proc. 1992 Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 279-285.
- [6] —, "Worst case identification in \mathcal{H}_{∞} : Validation of a priori information essentially optimal algorithms and error bounds," in *Proc. 1992 Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 251-257.
- [7] M. A. Dahleh, T. Theodosopoulos, and J. N. Tsitsiklis, "The sample complexity of worst case identification of FIR linear systems," *Syst. Contr. Lett.*, vol. 20, no. 3, Mar. 1993.
- [8] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input/Output Properties*. New York: Academic, 1975.
- [9] C. Foias and A. E. Frazho, *The Commutant Lifting Approach to Interpolation Problems*. Basel, Germany: Birkhäuser Verlag, 1990.
- [10] J. B. Garnett, *Bounded Analytic Functions*. New York: Academic, 1981.
- [11] G. Gu and P. P. Khargonekar, "Linear and nonlinear algorithms for identification in \mathcal{H}_{∞} with error bounds," *IEEE Trans. Automat. Contr.*, vol. 37, no. 7, pp. 953-963, July 1992.
- [12] —, "A class of algorithms for identification in \mathcal{H}_{∞} ," *Automatica*, vol. 28, pp. 229-312, Mar. 1992.
- [13] G. Gu, D. Xiong, and K. Zhou, "Identification in \mathcal{H}_{∞} using Pick's interpolation," in *Proc. 31st IEEE Conf. Decis. Contr.*, Tucson, AZ, pp. 1692-1693, Dec. 1992.
- [14] A. J. Helmicki, C. A. Jacobson, and C. N. Nett, "Control oriented system identification: A worst case/deterministic approach in \mathcal{H}_{∞} ," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1163-1176, Oct. 1991.
- [15] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge: Cambridge Univ. Press, 1985.
- [16] C. A. Jacobson, C. N. Nett, and J. R. Partington, "Worst case system identification in ℓ_1 : Optimal algorithms and error bounds," *Syst. Contr. Lett.*, vol. 19, pp. 419-424, 1992.
- [17] B. Kacwicz and M. Milanese, "Optimal finite sample experiment design in the worst-case ℓ_1 system identification," in *Proc. 31st IEEE Conf. Decis. Contr.*, Tucson, AZ, pp. 56-61, Dec. 1992.
- [18] M. G. Krein and A. A. Nudelman, *The Markov Moment Problem and Extremal Problems* (Translations of Math. Monographs), vol. 50. Providence, RI: American Math. Society, 1977.
- [19] N. Levinson and R. M. Redheffer, *Complex Variables*. New York: Holden Day, 1970.

- [20] P. M. Mäkilä, "Robust identification and Galois sequences," *Int. J. Contr.*, vol. 54, no. 5, pp. 1189–1200, 1991.
- [21] P. M. Mäkilä and J. R. Partington, "Robust identification of strongly stabilizable systems," *IEEE Trans. Automat. Contr.*, vol. 37, no. 11, pp. 1709–1716, Nov. 1992.
- [22] M. Milanese and R. Tempo, "Optimal algorithms theory for robust estimation and prediction," *IEEE Trans. Automat. Contr.*, vol. AC-30, no. 8, pp. 730–738, Aug. 1985.
- [23] M. Milanese and A. Vicino, "Optimal estimation theory for dynamic systems with set membership uncertainty: An overview," *Automatica*, vol. 27, no. 6, pp. 997–1009, 1991.
- [24] K. Y. Osipenko, "Optimal interpolation of analytic functions," *Matematicheskie Zametki*, vol. 12, no. 4, pp. 465–476, 1972.
- [25] M. Overton, "On minimizing the maximum eigenvalue of a symmetric matrix," *SIAM J. Matrix Anal. Appl.*, vol. 9, no. 2, pp. 256–268, 1988.
- [26] J. R. Partington, "Robust identification and interpolation in H_∞ ," *Int. J. Contr.*, vol. 54, pp. 1281–1290, 1991.
- [27] ———, "Robust identification in H_∞ ," *J. Math. Anal. and Appl.*, vol. 166, pp. 428–441, 1992.
- [28] K. Poolla, P. P. Khargonekar, A. Tikku, J. Krause, and K. M. Nagpal, "A time domain approach to model validation," in *Proc. 1992 Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 313–317.
- [29] K. Poolla and A. Tikku, "On the time complexity of worst-case system identification," in *Proc. 1993 Amer. Contr. Conf.*, San Francisco, CA, June 1993.
- [30] M. Rosenblum and J. Rovnyak, *Hardy Classes and Operator Theory*. New York: Oxford Univ. Press, 1985.
- [31] J. F. Traub, G. W. Wasilkowski, and H. Wozniakowski, *Information-Based Complexity*. New York: Academic, 1988.
- [32] D. C. N. Tse, M. A. Dahleh, and J. N. Tsitsiklis, "Optimal asymptotic identification under bounded disturbance," *IEEE Trans. Automat. Contr.*, vol. 38, no. 8, pp. 1176–1190, Aug. 1993.
- [33] T. Zhou and H. Kimura, "Time domain identification for robust control," *Syst. Contr. Lett.*, vol. 20, pp. 167–178, 1993.
- [34] Special Issue on System Identification for Robust Control Design, *IEEE Trans. Automat. Contr.*, vol. 37, no. 7, July 1992.

Pole Assignment for Linear Periodic Systems by Memoryless Output Feedback

Dirk Aeyels and Jacques L. Willems

Abstract—We consider linear periodic discrete-time systems. We are interested in the problem of placing the poles of the monodromy map by means of periodic output feedback of the same or multiple period. It is well known that, in general, the poles of time-invariant systems cannot be assigned by constant output feedback. This is in contrast with what can be obtained in the context of time-variant systems. The main contribution of this paper is that periodic output feedback suffices for pole placement of periodic systems.

I. STABILIZATION AND POLE ASSIGNMENT FOR LINEAR SYSTEMS

The purpose of closed-loop control is to monitor the behavior of a dynamical system by means of a well-chosen feedback to achieve a prescribed goal. The design of appropriate feedback laws to obtain stability or a preassigned location of the poles, as well as the study of the limitations imposed by constraints on the possible feedback strategies, represent major research areas in control theory. For

Manuscript received October 13, 1994. This work was supported in part by EC-Science Project SC1-0433-C(A).

The authors are with the Faculty of Engineering (Department of Systems Dynamics), Universiteit Gent, Technologiepark-Zwijnaarde, 9, B-9052 GENT (Zwijnaarde), Belgium.

IEEE Log Number 9408785.

linear time-invariant systems, the closed-loop poles can be assigned arbitrarily¹ by state feedback, provided the system is controllable. On the other hand, controllability and observability imply the existence of a state reconstructor, such that the poles of the closed-loop system can be assigned arbitrarily by dynamic output feedback.

It is well known, however, that in general pole assignment is not possible by means of linear time-invariant memoryless output feedback. In some recent papers [2], [3] the authors have shown that—under some technical conditions—for controllable and observable linear time-invariant discrete-time systems, exact pole assignment is possible by means of memoryless linear periodic output feedback. It seems natural to examine the possibilities of periodic feedback control when the original system itself is periodic. A first result, concerned with low-order systems, has appeared in [4]. In the present paper we extend these results to the general case of discrete-time systems. To our knowledge the problem has only received limited attention in the literature except for a recent paper that addresses the same question [7]. Our results are essentially different, e.g., there is no restriction on the dimension of the system with respect to the number of inputs and outputs.

II. MEMORYLESS PERIODIC OUTPUT FEEDBACK CONTROL FOR PERIODIC LINEAR SYSTEMS

A. Problem Statement

Consider a linear periodic discrete-time system with scalar input and scalar output

$$x_{i+1} = A_i x_i + b_i u_i, \quad y_i = c_i x_i \quad (1)$$

where $x_i \in R^n$, $u_i \in R$, $y_i \in R$, $i \in Z$. The matrices A_i , b_i , c_i have appropriate dimensions. The system is assumed to be periodic with period T

$$A_{i+T} = A_i, \quad b_{i+T} = b_i, \quad c_{i+T} = c_i.$$

The fundamental question that arises is to what extent the pole assignment problem can be solved by introducing time-varying (but memoryless) output feedback. In particular we assume that the time dependent feedback gain is periodic with the same period as the system. Consider the system (1) with periodic feedback

$$u_{i+T} = k_r y_{i+T}$$

with $r \in Z$ and $i \in \{0, 1, \dots, T-1\}$. With this periodic feedback the closed-loop system is also a periodic linear system of period T . This system can be considered as a time-invariant system over a time interval equal to the period T

$$\begin{aligned} x_{(i+1)T} &= (A_{T-1} + k_{T-1} b_{T-1} c_{T-1}) \\ &\times (A_{T-2} + k_{T-2} b_{T-2} c_{T-2}) \cdots (A_0 + k_0 b_0 c_0) x_{iT}. \end{aligned} \quad (2)$$

The eigenvalues of the system matrix

$$\begin{aligned} A_{cl} &= (A_{T-1} + k_{T-1} b_{T-1} c_{T-1}) \\ &\times (A_{T-2} + k_{T-2} b_{T-2} c_{T-2}) \cdots (A_0 + k_0 b_0 c_0) \end{aligned}$$

of (2)—from now on also called the poles of the periodic system—determine the dynamics of the periodic system. We recall here that the eigenvalues of A_{cl} are invariant under cyclic permutation

¹Obviously the complex poles must occur in complex conjugate pairs.

of its defining factors: this implies that the poles are identical for every system matrix that describes the closed-loop system considered over a period T . The original problem of pole assignment can then be restated as selecting feedback gains k_{T-1}, \dots, k_0 such that the eigenvalues of the closed-loop system matrix A_{cl} are the roots of a given polynomial:

$$\alpha(z) = z^n + \alpha_{n-1}z^{n-1} + \dots + \alpha_1z + \alpha_0. \quad (3)$$

B. Reformulation of the Problem

After introducing the notation

$$\Pi = (A_{T-2} + k_{T-2}b_{T-2}c_{T-2}) \cdots (A_0 + k_0b_0c_0) \\ A_{eq} = A_{T-1}\Pi, \quad c_{eq} = c_{T-1}\Pi, \quad b_{eq} = b_{T-1}$$

system (2) is represented as

$$\xi_{r+1} = A_{eq}\xi_r + b_{eq}v_r, \quad z_r = c_{eq}\xi_r \quad (4)$$

with $\xi_r = x_{r,T}$ and $v_r = k_{T-1}z_r$. The question is whether the eigenvalues of

$$A_{cl} = A_{eq} + k_{T-1}b_{eq}c_{eq} \quad (5)$$

can be arbitrarily assigned by the feedback gains. This question is treated in the Appendix. The analysis leads to the result stated and discussed in the next sections.

C. Main Result

The analysis of the appendixes yields the pole assignment property for systems with period T equal to $n+1$. Only this particular case is considered in the present section. The general case is referred to the discussion in Section II-D. For the sake of the presentation a number of assumptions are formulated before the theorem statement is given.

Assumption 1: The matrices A_i for $i = 0, \dots, n$, are nonsingular.

Assumption 2: The n vectors

$$\begin{bmatrix} c_1 \\ c_2 A_1 \\ \vdots \\ c_{n-1} A_{n-2} A_{n-3} \cdots A_1 \\ c_n A_{n-1} A_{n-2} \cdots A_1 \end{bmatrix}$$

are linearly independent.

Assumption 3: The n vectors

$$\begin{bmatrix} c_0 A_0^{-1} \\ c_1 \\ c_2 A_1 \\ \vdots \\ c_{n-1} A_{n-2} A_{n-3} \cdots A_1 \end{bmatrix}$$

are linearly independent.

Assumption 4:

$$\text{rank}[b_{eq} \quad A_{eq}b_{eq} \quad \cdots \quad A_{eq}^{n-1}b_{eq}] = n \quad (6)$$

for the values of the gains k_i , for $i = 1, \dots, n-1$, given by (18), and for k_0 equal to $k_0^0 = -1/c_0 A_0^{-1} b_0$.

Assumption 5:

$$\text{rank}[b_{eq} \quad A_{eq}b_{eq} \quad \cdots \quad A_{eq}^{n-2}b_{eq} \quad A_0^{-1}b_0] = n \quad (7)$$

for the values of the gains k_i , for $i = 1, \dots, n-1$, given by (18), and for k_0 equal to $k_0^0 = -1/c_0 A_0^{-1} b_0$.

Assumption 6: The parameters of system (1) satisfy

$$c_0 A_0^{-1} b_0 \neq 0.$$

The denominators of the gains k_i^0 , for $i = 1, \dots, n-1$, defined by (18), are nonzero

$$1 + k_i^0 c_i A_i^{-1} b_i \neq 0 \quad \text{for } i = 1, \dots, n-1.$$

Theorem 1: Consider the periodic single input single output system (1) with period $n+1$. Then under Assumptions 1)–6), the poles of the system can arbitrarily be assigned except at the origin by a periodic output feedback strategy with period equal to $n+1$.

D. Discussion

- A noteworthy aspect of Theorem 1 is that the feedback law has the same period $n+1$ as the system itself. Hence the system structure is not altered by the feedback, in contrast with the result for time-invariant systems [3], where the time-invariant system is transformed to a periodic system by the feedback. Systems with period T not equal to $n+1$ can readily be dealt with:

—If the period T is larger than $n+1$, then the result can still be used. Indeed if, e.g., at $rT+1, r \in \mathbb{Z}$, the input is set equal to zero, then

$$x_{rT+2} = A_1 A_0 x_{rT} + A_1 b_0 u_{rT}.$$

This reduces the period by one, considering $rT, rT+2, rT+3, \dots$ as sampling instants, $A_1 A_0$ as the new plant matrix, and $A_1 b_0$ as the new input matrix at rT . Repeating this procedure (if necessary), one finally obtains a system of period $n+1$ to which Theorem 2 can be applied. This periodic feedback actually $n+1$ corresponds to a feedback of period T in the original system.

Another approach would be to consider more variables k_i in the analysis of the appendixes, such that some of the variables can be chosen (almost) arbitrarily.

—If the period T is less than $n+1$, then the system can also be seen as a periodic system with period $2T, 3T$, or in general kT , with $k \in \mathbb{Z}$. In this way a period equal to $n+1$ or larger can easily be obtained, and the problem is reduced to either the case considered above, or the case considered in Theorem 1 itself.

- We briefly discuss the assumptions of the theorem. Assumption 1 on the nonsingularity of the system matrices is necessary for the analysis in the Appendix. This assumption, if not true, may be realized by an *a priori* output feedback; this feedback may be combined with the periodic output feedback obtained from the analysis of the Appendix resulting in a periodic output feedback that solves the pole assignment problem. If, however, Assumption 1 cannot be realized in this way, system (2) has a zero eigenvalue for all output feedback coefficients and pole assignment is certainly not possible by linear periodic output feedback. Assumptions 2–3 correspond to observability of the system. Assumptions 4–6 are weak conditions that are a consequence of the technical approach. It is also assumed that the desired pole pattern does not contain one or more closed-loop poles at the origin. This restriction is again a consequence of the technique and can most probably be waived. Some of the restrictions implied by the assumptions of Theorem 1 can be avoided by a cyclic permutation of the time indexes $0, 1, 2, \dots, T-1$. Also it can readily be seen that Assumptions 1–6 reduce to the assumptions obtained in [3] in case the original system is time-invariant.
- The result of Theorem 1 can be generalized to MIMO systems; see [3] for details.
- It seems important to emphasize that the result of Theorem 1 is more than a statement on the genericity of pole assignability by periodic memoryless output feedback. The theorem provides explicit sufficient conditions on the parameters under which the pole assignability problem is solvable. These conditions are

clearly generically satisfied since they require that the parameters do not satisfy a set of algebraic equations.

- The problem of pole assignment is basically one of solving a set of nonlinear algebraic equations. These equations are explicitly derived in Appendix B. We have given a number of verifiable conditions under which this problem is solvable.
- An important feature of the result of Theorem 1 is that no minimum number of inputs or outputs is imposed, distinguishing it from the result of Yan and Bitmead [7] who relate the number of inputs and outputs to the order of the system. Notice also that we are considering exact pole assignment, not almost pole assignment.

E. Illustrative Example

As an example, consider a periodic system of second order with period 3. The system data are

$$A_0 = \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix}, A_1 = \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix}$$

$$c_1 = c_2 = [-1 \quad 1]$$

We want to find output feedback gains such that for the system considered over one period the closed-loop poles are $-0.5 \pm 0.5j$. It is directly verified that the assumptions of the main theorem are satisfied. By the techniques explained in the appendixes the gains k_0 and k_1 must be determined from (13) and k_2 follows from Ackermann's formula. For this specific example (13) is

$$2(k_1 - 1)(k_0 - 3)(2k_1 k_0 + 4k_0 + 10) + (k_1 k_0 + k_0 - k_1 + 1) = 0.$$

The nominal choice for k_0 is here $k_0^0 = 3$. Following the perturbation analysis we choose $k_0 = 2$, and we obtain $k_1 = 1.0915$. From Ackermann's formula we obtain $k_2 = -3.4665$. This leads to

$$A_1 = [A_2 + k_2 b_2 c_2][A_1 + k_1 b_1 c_1][A_0 + k_0 b_0 c_0]$$

$$= \begin{bmatrix} 3.0915 & 6.0915 \\ -2.1585 & -1.0915 \end{bmatrix}$$

with the required eigenvalues.

III. CONCLUSION

After a discussion of the pole assignment problem for time-invariant discrete-time systems by memoryless output feedback we have posed the same problem for periodic systems. The main part of the paper is concerned with establishing the following result: it is possible in general to assign the poles of a linear periodic discrete-time system by linear periodic output feedback of the same or multiple period.

APPENDIX A

VAN DER WOUDE'S LEMMA

In Section II-B it was shown that the problem is reduced to finding the feedback gains such that the eigenvalues of (5) are the roots of a given characteristic polynomial (3). A result derived by Van der Woude [6] is well suited to analyze this condition. It is formulated below applied to the problem formulated in Section II-B.

As a general rule we will indicate in the course of the technical developments where the assumptions of the main theorem come in. This is done by explicitly making reference to them.

Lemma 1: If the pair $A_{i,q}, b_{i,q}$ is controllable (Assumption 4), there exists an output feedback

(8)

for system (4) such that the polynomial $\alpha(\cdot)$, given by (3), is the characteristic polynomial of (5), if and only if

$$\alpha(A_{i,q})\ker(c_{i,q}) \subset \text{im}[b_{i,q} \quad A_{i,q}b_{i,q} \quad \dots \quad A_{i,q}^{T-1}b_{i,q}] \quad (9)$$

The problem is hence to ensure, if possible, that (9) can be satisfied by a suitable choice of real k_{T-2}, \dots, k_0 . Notice that both sides of (9) depend on k_{T-2}, \dots, k_0 through A_{eq} and c_{eq} . Once k_{T-2}, \dots, k_0 have been computed, it must be checked that the pair A_{eq}, b_{eq} is controllable. The control, given by (8), must then be determined such that

$$\alpha(\cdot) = \det[zI - (A_{eq} - k_{T-1}b_{eq}c_{eq})].$$

The existence of the control follows from Lemma 1. One approach for determining k_{T-1} is to use Ackermann's formula [1], [5]. The original problem is then reduced to ensure that the feedback constants can be chosen such that (9) holds and such that the controllability matrix has full rank

$$\text{rank}[b_{eq} \quad A_{eq}b_{eq} \quad \dots \quad A_{eq}^{T-1}b_{eq}] \quad (10)$$

The following appendixes are dedicated to an in depth analysis of relations (9) and (10). The main line of the analysis is as follows. First (9) is reformulated as an algebraic equation in k_{T-2}, \dots, k_0 . This equation is explicitly solved in Appendix B for a special choice of k_0 which is denoted by k_0^0 ; the obtained solutions are denoted as k_{T-2}^0, \dots, k_1^0 . These solutions are not acceptable, however, since the choice of k_0 violates one of the conditions imposed. Therefore a perturbation analysis is performed in Appendix C to show that for k_0 close to k_0^0 the equation can also be solved and the solutions can be accepted. The analysis requires some weak restrictions on the system data, which are introduced as the discussion proceeds.

APPENDIX B

THE DETERMINATION OF THE NOMINAL FEEDBACK GAINS

In the analysis of the present and the next appendix, the period T is set equal to $n+1$. This is in accordance with the results obtained for the time-invariant case [3]. In the sequel we assume that all matrices A_i are nonsingular (Assumption 1).

To solve (9), the equation is transformed into a more tractable expression under the additional assumption that the matrices $A_i + k_i b_i c_i$, for $i = 0, \dots, n-1$, are nonsingular. Then

$$\ker(c_{i,q}) = \ker(c_n \Pi) = \Pi^{-1} \ker(c_n) \quad (11)$$

and (9) is equivalent with

$$\alpha(A_{i,q})\Pi^{-1} \ker(c_n) \subset \text{im}[b_n \quad A_n \Pi b_n \quad \dots \quad (A_n \Pi)^{n-2} b_n] \quad (12)$$

Of course it must be checked *a posteriori* that the nonsingularity conditions are satisfied for the obtained solutions.

Assume that the controllability condition (10) is satisfied and let v^+ be a row vector (dependent on k_i) perpendicular to the $n-1$ independent vectors

$$b_n, A_n \Pi b_n, (A_n \Pi)^2 b_n, \dots, (A_n \Pi)^{n-2} b_n.$$

Let $c_1^+, c_2^+, \dots, c_{n-1}^+$ be a basis of column vectors spanning the null space of c_n (i.e., perpendicular to c_n). Then (12) is equivalent with the set of $n-1$ equations in the n variables k_0, k_1, \dots, k_{n-1}

$$v^+ \alpha(A_{eq}) \Pi^{-1} c_i^+ = 0$$

for $i = 1, \dots, n-1$. Explicitly these equations are

$$v^+ (A_{eq}^{n-1} A_n + \dots + \alpha_1 A_n) c_i^+ + \alpha_0 v^+ (A_0 + k_0 b_0 c_0)^{-1} \dots (A_{n-1} + k_{n-1} b_{n-1} c_{n-1})^{-1} c_i^+ = 0. \quad (13)$$

The Woodbury formula [5] yields

$$(A + kbc)^{-1} = (1 + kcA^{-1}b)^{-1}[(1 + kcA^{-1}b)A^{-1} - kA^{-1}bcA^{-1}].$$

Then (13) is equivalent with

$$\begin{aligned} & (1 + k_0 c_0 A_0^{-1} b_0) v^+ \\ & \times (A_{n-1}^{-1} A_n + \alpha_{n-1} A_{n-1}^{-2} A_n + \cdots + \alpha_1 A_n) c_i^+ \\ & + \alpha_0 v^+ [(1 + k_0 c_0 A_0^{-1} b_0) A_0^{-1} - k_0 A_0^{-1} b_0 c_0 A_0^{-1}] \\ & \times (A_1 + k_1 b_1 c_1)^{-1} \cdots (A_{n-1} + k_{n-1} b_{n-1} c_{n-1})^{-1} c_i^+ = 0 \end{aligned} \quad (14)$$

for $i = 1, \dots, n-1$. Recall that the desired gains k_i must be such that the matrices $A_i + k_i b_i c_i$, with $i = 0, \dots, n-1$, are nonsingular. Note that (14) consists of $n-1$ equations with n variables k_i ; hence in general one of the gains may be chosen and the others computed from the equations. We assume that $c_0 A_0^{-1} b_0$ does not vanish (Assumption 6). Although we have required that the matrices $A_i + k_i b_i c_i$ are nonsingular, we first solve (14) under the assumption that k_0 is equal to $-1/c_0 A_0^{-1} b_0$, denoted as k_0^0 . Appendix C is devoted to solving (14) for k_0 different from that value.

The gains k_i satisfying (14) for the particular choice of k_0 are called the nominal solutions; they are denoted by k_i^0 . It is assumed that α_0 is nonzero. This implies that the desired closed-loop system has no poles at the origin. Furthermore we assume that $v^+ A_0^{-1} b_0 \neq 0$. This condition holds due to Assumptions 4 and 5. Then (14) yields

$$c_0 A_0^{-1} (A_1 + k_1 b_1 c_1)^{-1} \cdots (A_{n-1} + k_{n-1} b_{n-1} c_{n-1})^{-1} v^+ = 0. \quad (15)$$

This is equivalent with

$$c_0 A_0^{-1} \div c_n (A_{n-1} + k_{n-1} b_{n-1} c_{n-1}) \cdots (A_1 + k_1 b_1 c_1) \quad (16)$$

where the symbol \div stands for proportionality.

Although the above set of equations for the gains k_1, \dots, k_{n-1} is nonlinear, it is shown below that it can be brought in a triangular form and explicitly solved. Assume that the vectors

$$v_n = c_n A_{n-1} A_{n-2} \cdots A_1$$

are linearly independent (Assumption 2). This is actually an observability condition, requiring that the state can be reconstructed from n successive output measurements. Then $c_0 A_0^{-1}$ can be expressed as

$$c_0 A_0^{-1} = a_1 v_1 + a_2 v_2 + \cdots + a_n v_n, \quad (17)$$

The constant a_n is assumed to be nonzero (this is Assumption 3); it determines the proportionality factor in (16). The solutions k_i for $i = 1, \dots, n-1$ can readily be calculated from (16); the solution k_i^0 is expressed as a function of k_{i+1}^0, \dots, k_n^0 :

$$k_i^0 = \frac{a_i}{a_n c_n (A_{n-1} + k_{n-1}^0 b_{n-1} c_{n-1}) \cdots (A_{i+1} + k_{i+1}^0 b_{i+1} c_{i+1}) b_i},$$

where the denominators are assumed to be nonzero. More explicitly for $i = 1, \dots, n-1$

$$k_i^0 = \frac{a_i}{a_n c_n A_{n-1} \cdots A_{i+1} b_i + \cdots + a_{i+1} c_{i+1} b_i}, \quad (18)$$

APPENDIX C PERTURBATION ANALYSIS

In Appendix B we obtained a set of expressions (18) determining the constants k_i^0 ($i = 1, \dots, n$) satisfying relation (12) for a choice of k_0 equal to k_0^0 . This value cannot be accepted as was explained in Appendix B. Therefore, relation (12) is reconsidered for $k_0 = k_0^0 + \epsilon$, with ϵ small. It is shown below that for this choice of k_0 there exist solutions k_i ($i = 1, \dots, n-1$) of (12) or equivalently (14).

The application of the Woodbury formula to all matrices $(A_i + k_i b_i c_i)^{-1}$ in (14) yields

$$\begin{aligned} & (1 + k_0 c_0 A_0^{-1} b_0) (1 + k_1 c_1 A_1^{-1} b_1) \\ & \cdots (1 + k_{n-1} c_{n-1} A_{n-1}^{-1} b_{n-1}) \\ & \times v^+ (A_{n-1}^{-1} A_n + \alpha_{n-1} A_{n-1}^{-2} A_n + \cdots + \alpha_1 A_n) c_i^+ \\ & + \alpha_0 v^+ [(1 + k_0 c_0 A_0^{-1} b_0) A_0^{-1} - k_0 A_0^{-1} b_0 c_0 A_0^{-1}] \\ & \times [(1 + k_{n-1} c_{n-1} A_{n-1}^{-1} b_{n-1}) A_{n-1}^{-1} \\ & - k_{n-1} A_{n-1}^{-1} b_{n-1} c_{n-1} A_{n-1}^{-1}] c_i^+ = 0 \end{aligned} \quad (19)$$

for $i = 1, \dots, n-1$. The vectors c_i^+ constitute a basis of $\ker(c_n)$. We take the following particular choice. Consider the set of vectors

$$w_i = v_i A_1^{-1} A_2^{-1} \cdots A_{n-1}^{-1}$$

for $i = 1, \dots, n$. These vectors are independent because of the observability assumption introduced in Appendix B. Notice that

$$w_i = v_i A_1^{-1} A_2^{-1} \cdots A_{n-1}^{-1} = v_i A_i^{-1} A_{i+1}^{-1} \cdots A_{n-1}^{-1}$$

for $i = 1, \dots, n-1$, and

The vectors c_i^+ are chosen such that c_i^+ is orthogonal to all w_j , for $j \neq i$. Each vector is unambiguously defined up to a scalar.

In an obvious notation each equation in (19) is written as

$$\Phi_i(k_0, \dots, k_{n-1}) + \alpha_0 g_i(k_0, \dots, k_{n-1}) = 0 \quad (20)$$

for $i = 1, \dots, n-1$. Let k^0 denote the vector of the nominal gains k_0^0, \dots, k_{n-1}^0 . Below it is shown that the Jacobian matrix

$$\frac{\partial(\Phi_i + \alpha_0 g_i)}{\partial k_j} \quad (21)$$

with $i, j = 1, \dots, n-1$, evaluated at k^0 is nonsingular. This establishes by the implicit function theorem that there exists a solution to (19) also for $k_0 = k_0^0 + \epsilon$ with ϵ sufficiently small. The corresponding solutions k_i are smoothly dependent on ϵ and equal to k_i^0 for ϵ equal to zero. Assume (Assumption 6)

$$k_i^0 c_i A_i^{-1} b_i \neq -1 \quad \text{for } i = 1, \dots, n-1. \quad (22)$$

Then the solutions k_i also satisfy $1 + k_i c_i A_i^{-1} b_i \neq 0$ for small ϵ .

We now prove the nonsingularity of the Jacobian matrix. First notice that evaluated at k^0 , we have

$$\frac{\partial(\Phi_i + \alpha_0 g_i)}{\partial k_j} = \alpha_0 \frac{\partial g_i}{\partial k_j} \quad (23)$$

for $i, j = 1, \dots, n-1$, since each Φ_i contains the factor $1 + k_0 c_0 A_0^{-1} b_0$ which vanishes at k^0 . For $k_0 = k_0^0$ the function g_i can be written as follows

$$\frac{1}{c_0 A_0^{-1} b_0} f_i$$

with

$$\begin{aligned} f_i &= c_0 A_0^{-1} [(1 + k_1 c_1 A_1^{-1} b_1) A_1^{-1} - k_1 A_1^{-1} b_1 c_1 A_1^{-1}] \\ & \times [(1 + k_{n-1} c_{n-1} A_{n-1}^{-1} b_{n-1}) A_{n-1}^{-1} \\ & - k_{n-1} A_{n-1}^{-1} b_{n-1} c_{n-1} A_{n-1}^{-1}] c_i^+, \end{aligned}$$

From (15) it follows that f_i vanishes at k^0 . Therefore

$$\alpha_0 \frac{\partial f_i}{\partial k_j} = \alpha_0 \frac{v^+ A_0^{-1} b_0}{c_0 A_0^{-1} b_0} \frac{\partial f_i}{\partial k_j}. \quad (24)$$

Notice that with the particular choice of the vectors c_1^+ the function f_1 has the following form

$$\begin{aligned} f_1 &= (1 + k_2 c_2 A_2^{-1} b_2) \cdots (1 + k_{n-1} c_{n-1} A_{n-1}^{-1} b_{n-1}) \\ &\quad \times c_0 A_0^{-1} [A_1^{-1} (1 + k_1 c_1 A_1^{-1} b_1) \\ &\quad - k_1 A_1^{-1} b_1 c_1 A_1^{-1}] A_2^{-1} A_3^{-1} \cdots A_{n-2}^{-1} A_{n-1}^{-1} c_1^+. \end{aligned}$$

With the particular choice of c_1^+ and invoking (17) one obtains

$$c_0 A_0^{-1} A_1^{-1} \cdots A_{n-1}^{-1} c_1^+ = a_1 c_1 A_1^{-1} A_2^{-1} \cdots A_{n-1}^{-1} c_1^+.$$

Hence

$$\begin{aligned} f_1 &= (1 + k_2 c_2 A_2^{-1} b_2) \cdots (1 + k_{n-1} c_{n-1} A_{n-1}^{-1} b_{n-1}) \\ &\quad \times [a_1 (1 + k_1 c_1 A_1^{-1} b_1) - k_1 c_0 A_0^{-1} A_1^{-1} b_1] \\ &\quad \times c_1 A_1^{-1} \cdots A_{n-1}^{-1} c_1^+. \end{aligned}$$

With

$$c_0 A_0^{-1} = a_1 c_1 + a_2 c_2 A_1 + \cdots + a_n c_n A_{n-1} \cdots A_1 \quad (25)$$

one obtains

$$\begin{aligned} f_1 &= (1 + k_2 c_2 A_2^{-1} b_2) \cdots (1 + k_{n-1} c_{n-1} A_{n-1}^{-1} b_{n-1}) \\ &\quad \times c_1 A_1^{-1} \cdots A_{n-1}^{-1} c_1^+ [a_1 - k_1 (a_2 c_2 + a_3 c_3 A_2 \\ &\quad + \cdots + a_n c_n A_{n-1} A_{n-2} \cdots A_2) b_1]. \end{aligned}$$

From the explicit expressions (18) of k_i^0 we conclude that all elements (except for the diagonal element) of the first row of the Jacobian matrix (21) vanish when they are evaluated at k^0

$$\frac{\partial f_1}{\partial k_j}(k^0) = 0$$

for $j > 1$.

Consider the first diagonal element: one obtains

$$\begin{aligned} \frac{\partial f_1}{\partial k_1}(k^0) &= -(1 + k_2^0 c_2 A_2^{-1} b_2) \cdots (1 + k_{n-1}^0 c_{n-1} A_{n-1}^{-1} b_{n-1}) \\ &\quad \times c_1 A_1^{-1} \cdots A_{n-1}^{-1} c_1^+ (a_2 c_2 + a_3 c_3 A_2 \\ &\quad + \cdots + a_n c_n A_{n-1} A_{n-2} \cdots A_2) b_1. \end{aligned}$$

This expression is different from zero since (22) holds and the nominal k_i^0 exist (i.e., the denominators are not equal to zero).

Consider now the second row of the Jacobian matrix (21): the expression of f_2 is rewritten, using the definition of c_2^+

$$\begin{aligned} f_2 &= (1 + k_1 c_1 A_1^{-1} b_1) \cdots (1 + k_{n-1} c_{n-1} A_{n-1}^{-1} b_{n-1}) \\ &\quad \times c_0 A_0^{-1} [(1 + k_1 c_1 A_1^{-1} b_1) A_1^{-1} - k_1 A_1^{-1} b_1 c_1 A_1^{-1}] \\ &\quad \times [(1 + k_2 c_2 A_2^{-1} b_2) A_2^{-1} - k_2 A_2^{-1} b_2 c_2 A_2^{-1}] \\ &\quad \times A_3^{-1} \cdots A_{n-2}^{-1} A_{n-1}^{-1} c_2^+. \end{aligned}$$

Invoking (17) and the definition of c_2^+ , one obtains

$$\begin{aligned} f_2 &= (1 + k_1 c_1 A_1^{-1} b_1) \cdots (1 + k_{n-1} c_{n-1} A_{n-1}^{-1} b_{n-1}) \\ &\quad \times [(1 + k_1 c_1 A_1^{-1} b_1) (1 + k_2 c_2 A_2^{-1} b_2) a_2 \\ &\quad - (1 + k_1 c_1 A_1^{-1} b_1) k_2 c_0 A_0^{-1} A_1^{-1} A_2^{-1} b_2 \\ &\quad + k_1 k_2 c_0 A_0^{-1} A_1^{-1} b_1 c_1 A_1^{-1} A_2^{-1} b_2] c_2 \\ &\quad \times A_3^{-1} \cdots A_{n-2}^{-1} A_{n-1}^{-1} c_2^+. \end{aligned}$$

One immediately verifies that

$$\frac{\partial f_2}{\partial k_j}(k^0) = 0$$

for $j > 2$. In general it may be verified that

$$\frac{\partial f_i}{\partial k_j}(k^0) = 0$$

for $j > i$.

As for the second diagonal element of the Jacobian matrix one obtains

$$\begin{aligned} \frac{\partial f_2}{\partial k_2}(k^0) &= -(1 + k_1^0 c_1 A_1^{-1} b_1) (1 + k_3^0 c_3 A_3^{-1} b_3) \\ &\quad \cdots (1 + k_{n-1}^0 c_{n-1} A_{n-1}^{-1} b_{n-1}) \\ &\quad \times c_2 A_2^{-1} \cdots A_{n-1}^{-1} c_2^+ (a_3 c_3 + a_4 c_4 A_3 \\ &\quad + \cdots + a_n c_n A_{n-1} A_{n-2} \cdots A_3) b_2. \end{aligned}$$

This expression is different from zero since (22) holds and the nominal k_i^0 exist (i.e., the denominators are not equal to zero).

Similar conclusions are obtained for the other diagonal elements of the Jacobian matrix. Therefore the Jacobian matrix, evaluated at k^0 , is nonsingular if

- (Assumption 6) the nominal gains k_i^0 , for $i = 1, \dots, n-1$, exist and satisfy (22),
- $v^+ A_0^{-1} b_0 \neq 0$, $c_0 A_0^{-1} b_0 \neq 0$ (Assumption 6) and $\alpha_0 \neq 0$.

Note that $v^+ A_0^{-1} b_0 \neq 0$ is equivalent to (Assumption 5)

$$\text{rank}[b_{vq} \quad A_{1q} b_{vq} \quad \cdots \quad A_{n-1,q}^{n-2} b_{vq} \quad A_0^{-1} b_0] = n \quad (26)$$

if the controllability condition (10) is satisfied at k^0 .

ACKNOWLEDGMENT

This paper presents results of the Belgian Programme on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office for Science, Technology, and Culture. The scientific responsibility rests with its authors.

REFERENCES

- [1] J. Ackermann, *Abtastregelung*. Berlin: Springer-Verlag, 1972.
- [2] D. Aeyels and J. L. Willems, "Pole assignment for linear time-invariant second-order systems by static output feedback," *IMA J. Math. Contr. Inform.*, vol. 8, pp. 267-274, 1991.
- [3] —, "Pole assignment for linear time-invariant systems by periodic memoryless output feedback," *Automatica*, vol. 28, pp. 1159-1168, 1992.
- [4] —, "Pole assignment for linear periodic systems by means of periodic memoryless output feedback," in *Proc. 32nd Conf. Decis. Contr.*, San Antonio, TX, 1993, pp. 1361-1362.
- [5] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1980.
- [6] J. W. Van der Woude, "A note on pole placement by static output feedback for single-input systems," *Syst. Contr. Lett.*, vol. 11, pp. 285-287, 1988.
- [7] W. Y. Yan and R. R. Bitmead, "Control of linear discrete-time periodic systems: a decentralized control approach," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1644-1648, 1992.

Fault Detection and Isolation for Unstable Linear Systems

M. Kinnaert, R. Hanus, and Ph. Arte

Abstract—In the design of a residual generator for an unstable plant, the unstable modes are made unobservable from the residual vector. We point out that this cannot be achieved in practice due to the discrepancy between the estimated model and the “true” plant model. Hence, we deduce that it is not possible to design a residual generator for an unstable plant, unless it is stabilized by an adequate controller. As a by-product, we provide expressions which can be used to quantify the effect of modeling uncertainties on the performance of a residual generator. The factorization approach to control theory is used for the developments.

I. INTRODUCTION

There are many approaches to perform the synthesis of a residual generator for automatic fault detection on linear time invariant systems: the geometric approach [1]–[2], the factorization approach [3]–[5], the parity space approach [6]–[8], and the approach based on eigenstructure assignment [9]–[10]. The stability of the plant never appears as a necessary condition for the design of a residual generator in these works. We show here that this is due to the assumption that the estimated plant model exactly coincides with the “true” plant model. By introducing a mismatch between both models, we prove that it is not possible to design a residual generator for unstable plants when they are not stabilized by an adequate controller. Beside its theoretical interest, this remark might be useful when one considers unstable processes running during a finite time period or processes containing an integrating action.

This note is organized as follows. In the second part, the so called extension of the fundamental problem of residual generation (EFPRG) [1] is stated. In Section III, its solution using the factorization approach is reviewed, and the effect of the modeling error is exhibited. In the fourth section, the EFPRG is solved for an unstable plant controlled by a stabilizing controller, and the influence of modeling uncertainties is analyzed.

II. THE EXTENSION OF THE FUNDAMENTAL PROBLEM OF RESIDUAL GENERATION

For the sake of simplicity and to clearly show the influence of the initial state of the system, we first state the problem in the time domain.

We consider the class of systems described by

$$\dot{x}(t) = Ax(t) + Bu(t) + Ff(t) \quad (1)$$

$$y(t) = Cx(t) + Du(t) + Gf(t) \quad (2)$$

where $x(t) \in R^n$ is the state vector, $u(t) \in R^r$ is the actuator command vector, $y(t) \in R^l$ is the measurement vector, and $f(t) \in R^q$ is a vector made of arbitrary functions of time referred to as the failure modes.

Our aim is to design a linear time invariant (LTI) system with inputs $u(t)$ and $y(t)$, and with q outputs, namely the q components of a vector $r(t)$, called the residual vector, so that the map from $f(t)$ to $r(t)$ fulfills specific requirements. More precisely, a nonzero i th component of $f(t)$, $f_i(t)$, must induce a nonzero i th component of $r(t)$, $r_i(t)$, and it cannot influence the other components of $r(t)$. The

most general form for such a LTI system is

$$\dot{x}_i(t) = A_i x_i(t) + B_i u(t) + \Gamma_i y(t) \quad (3)$$

$$r_i(t) = C_i x_i(t) + D_i u(t) + F_i y(t) \quad (4)$$

Combining (1)–(4), we obtain

$$\begin{bmatrix} \dot{x}_i(t) \\ r_i(t) \end{bmatrix} = \begin{bmatrix} A_i & 0 \\ F_i C_i & A_i \end{bmatrix} \begin{bmatrix} x_i(t) \\ r_i(t) \end{bmatrix} + \begin{bmatrix} B_i \\ B_i + E_i D_i \end{bmatrix} u(t) + \begin{bmatrix} \Gamma_i \\ F_i \Gamma_i \end{bmatrix} f(t) \quad (5)$$

$$\begin{bmatrix} \dot{x}_i(t) \\ r_i(t) \end{bmatrix} = \begin{bmatrix} \Gamma_i C_i & C_i \end{bmatrix} \begin{bmatrix} x_i(t) \\ r_i(t) \end{bmatrix} + (F_i D_i + D_i) u(t) + F_i \Gamma_i f(t) \quad (6)$$

The objectives can now be stated as follows:

- O1) In the absence of failure, $r_i(t)$ decays asymptotically to zero $i = 1, 2, \dots, q$.
- O2) In the presence of the i th failure mode ($f_i(t) \neq 0$), $r_i(t)$ decays asymptotically to zero $j = 1, 2, \dots, i-1, i+1, \dots, q$ and $r_i(t)$ depends on $f_i(t)$ in a sense made precise below in R2). This should be verified for $i = 1, 2, \dots, q$.

This yields the following requirements on the system described by (5) and (6):

- R1) All the observable modes of the pair $\left(\begin{bmatrix} F_i C_i & C_i \end{bmatrix}, \begin{bmatrix} A_i & 0 \\ \Gamma_i C_i & A_i \end{bmatrix} \right)$ are asymptotically stable and the map from $u(t)$ to $r_i(t)$ is equal to zero.
- R2) The map from $f_i(t)$ to $r_i(t)$ corresponds to a diagonal transfer matrix with nonzero diagonal elements.

Remark 1

—Although the failure signals $f_i(t)$, $i = 1, \dots, q$, are arbitrary, we do not consider the situation where the system initial condition and the signal $f_i(t) \neq 0$ are such that $r_i(t) = 0$ due to a transmission zero. The probability of the occurrence of this phenomenon is zero for all practical purpose.

The observability condition in R1) guarantees that the effect on $r_i(t)$ of a nonzero initial state of the process or the filter asymptotically vanishes.

—The EFPRG is stated in a more general framework in [1]. Indeed, a failure can be modeled by several nonzero components in $f(t)$. In most practical situations, however, a single nonzero entry suffices to model a failure. Hence, we only consider this situation here.

—The diagonal fault response (i.e., hypothesis R2)) is a particular case. Yet, any other kind of response can be derived from it by an additional transformation [8].

It is now straightforward to translate the problem in the frequency domain. Taking into account the transient due to a nonzero initial state x^0 , the Laplace transform of (1) and (2) yields

$$y(s) = G(s)u(s) + G_f(s)f(s) + G_0(s)x^0 \quad (7)$$

where $G(s) = C(sI - A)^{-1}B + D$, $G_f(s) = C(sI - A)^{-1}F + F$, and $G_0(s) = C(sI - A)^{-1}x^0$.

Now, as only the observable part of the filter is useful, we assume without loss of generality that (C, A) is observable. Hence, to fulfill R1), A must be a Hurwitz matrix. Thus, we can take the Laplace transform of (3)–(4) and neglect the transient due to the initial state of the filter, as it decays asymptotically to zero. This yields

$$r_i(s) = L_i(s)y(s) + K_i(s)u(s) \quad (8)$$

Manuscript received April 26, 1994.

The authors are with the Laboratoire d'Automatique, Université Libre de Bruxelles, CP165 ULB, 50 Ave F. D. Roosevelt, B-1050, Brussels, Belgium. IEEE Log Number 9408786.

where $K(s) = C_r(sI - A_r)^{-1}B_r + D_r$, $L(s) = C_r(sI - A_r)^{-1}E_r + F_r$, and $L(s), K(s) \in RH_\infty$. Here, and in the sequel, RH_∞ denotes the set of stable proper rational matrices.

Substituting $y(s)$ by (7) in (8) yields

$$r(s) = (L(s)G(s) + K(s))u(s) + L(s)G_f(s)f(s) + L(s)G_0(s)x^0. \quad (9)$$

From (9), we deduce that R1) and R2) are equivalent to the following.

R1') $L(s)G(s) + K(s) = 0$ $L(s), K(s)$ and $L(s)G_0(s)$ belong to RH_∞ .

R2') $L(s)G_f(s)$ is a diagonal transfer matrix with nonzero diagonal entries.

III. EFFECT OF UNSTABLE MODES IN THE PLANT

We first review the class of all the solutions to EFPRG. For the sake of simplicity, we assume that the number of measurements, p , is the same as the number of failures, q . The more general case where $p > q$ can be handled as in [5].

The following lemma is stated slightly differently in [5]. Indeed, here, to simplify the equations, we have used the fact that there exist matrices $N(s)$, $N_f(s)$, and $N_0(s)$ such that

$$G(s) = M(s)^{-1}N(s) \quad (10a)$$

$$G_f(s) = M(s)^{-1}N_f(s) \quad (10b)$$

$$G_0(s) = M(s)^{-1}N_0(s) \quad (10c)$$

are left coprime factorizations of $G(s)$, $G_f(s)$, and $G_0(s)$ over RH_∞ .

Lemma [5]: Assume that $G_f(s)$ is invertible, and let (10a)–(10b) be left coprime factorizations of $G(s)$ and $G_f(s)$ over RH_∞ . Then the class of all stable proper filters which solve the extension of the fundamental problem of residual generation (i.e., which fulfill R1) and R2)), for the system (7), is parameterized as follows

$$r(s) = P_{N_f}(s)(M(s)y(s) - N(s)u(s))$$

where

$$P_{N_f}(s) = Q(s)S_{N_f}(s)N_f(s)^{-1} \quad (12)$$

with $Q(s)$, an arbitrary $q \times q$ diagonal RH_∞ matrix with nonzero diagonal entries

$$S_{N_f}(s) = \text{diag} \frac{\prod_{j=1}^{n_i} (s + z_j)}{(s + \alpha_i)^{n_i + p_i}} = \text{diag } s_i(s) \quad (13)$$

where z_j , $j = 1, \dots, n_i$ are the unstable poles of the i th row of $N_f(s)^{-1}$, α_i is an arbitrary positive real number and p_i is determined so that

$$\lim_{s \rightarrow \infty} s_i(s)[N_f(s)]^{-1} \quad (14)$$

is a nonzero row vector. Here $[X]_i$ denotes the i th row of X .

To show what goes wrong with an unstable plant, we now introduce a discrepancy between the "true" plant model (7), and the estimated plant model, denoted by hats ("^"), which is used to design the residual generator.

In this framework, (9) can be written

$$r(s) = (\hat{L}(s)G(s) + \hat{K}(s))u(s) + \hat{L}(s)G_f(s)f(s) + \hat{L}(s)G_0(s)x^0 \quad (15)$$

where

$$\hat{L}(s) = P_{\hat{N}_f}(s)\hat{M}(s); \quad \hat{K}(s) = -P_{\hat{N}_f}(s)\hat{N}(s). \quad (16)$$

Here $P_{\hat{N}_f}(s)$ is obtained by substituting $N_f(s)$ by $\hat{N}_f(s)$ in (12)–(14).

Now, using (10c), one can rewrite the map from x^0 to r as

$$\hat{L}(s)G_0(s) = P_{\hat{N}_f}(s)\hat{M}(s)M(s)^{-1}N_0(s). \quad (17)$$

From (17), one concludes that, due to the modeling error, when the plant is unstable, the effect of the initial conditions will not vanish. On the contrary a nonzero x^0 will yield an "explosion" in the residual. From (15)–(17) it is clear that whatever $\hat{L}(s)$ may be, it is impossible to avoid instability when $\hat{M}(s) \neq M(s)$. Notice that it is also impossible to assure that the effect of the initial condition on the residual vanishes for a plant with a pole at the origin. The residual will be affected by a bias whose magnitude depends on the initial state, in this case.

We now investigate the situation where a controller stabilizes the unstable plant.

IV. THE EFPRG FOR AN UNSTABLE PROCESS STABILIZED IN CLOSED LOOP

Let us consider a stabilizing controller

$$u(s) = C(s)(w(s) - y(s)). \quad (18)$$

For the plant (7), and let $N_r(s)M_r(s)^{-1}$ be a right coprime factorization of $C(s)$ over RH_∞ . In (18), $w(s)$ is the Laplace transform of the reference vector.

We can now state the following theorem.

Theorem: Consider the stable closed-loop system (7), (18), and assume that $G_f(s)$ is invertible. Let $\hat{C}(s) = \hat{M}(s)^{-1}\hat{N}(s)$, $\hat{G}_f(s) = \hat{M}(s)^{-1}\hat{N}_f(s)$, and $\hat{G}_0(s) = \hat{M}(s)^{-1}\hat{N}_0(s)$ be left coprime factorizations of the transfer matrices of the estimated plant model. Then the class of filters parametrized by

$$r(s) = \hat{L}(s)y(s) + \hat{K}(s)u(s) \quad (19)$$

where $\hat{L}(s)$ and $\hat{K}(s)$ are given in (16) fulfills

$$\lim_{s \rightarrow \infty} sr(s) = \Gamma w_0 \quad (20)$$

in the absence of failure, for a step-like reference signal $w(s) = w_0/s$, $w_0 \in R^p$. Here Γ is a constant real matrix.

Moreover, in the presence of step-like failure $f(s) = f_0/s$, $f_0 \in R^q$, and reference $w(s) = w_0/s$

$$\lim_{s \rightarrow \infty} sr(s) = \Gamma w_0 + \Theta f_0 \quad (21)$$

where Θ is a constant real matrix. Besides, when \hat{G} and \hat{G}_f , respectively, tend to G and G_f , Γ tends to zero, and Θ tends to a diagonal matrix.

Proof: In the sequel, the Laplace variable s is omitted to shorten the expressions. To find an expression of $r(s)$ in terms of $w(s)$, $f(s)$, and x^0 , we first write down the equations of the closed-loop systems (7) and (18), where we substitute the different transfer functions by their factorization. This yields

$$y = M_c D_{cl}^{-1} N_f f + M_c D_{cl}^{-1} N N_r M_c^{-1} w + M_c D_{cl}^{-1} N_0 x^0 \quad (22)$$

where $D_{cl} = MM_c + NN_r$. As the controller assures the internal stability of the closed-loop system by hypothesis, D_{cl} is a unimodular matrix [11]. Substituting $y(s)$ and $u(s)$ by (22) and (18), respectively, in (19) of the residual generator, we obtain

$$r = (\hat{L}M_c - \hat{K}N_r)D_{cl}^{-1}N_f f + (\hat{L}M_c - \hat{K}N_r)D_{cl}^{-1}N_0 x^0 + (\hat{L}M_c D_{cl}^{-1}NN_r M_c^{-1} + \hat{K}N_r D_{cl}^{-1}M)w. \quad (23)$$

As $\hat{L}(s)$ and $\hat{K}(s)$ are asymptotically stable, so are the transfer matrices between $f(s)$, x^0 , $w(s)$, and $r(s)$. This is obvious, except for the map from $w(s)$ to $r(s)$. Notice, however, that

$$\hat{L}M_c D_{cl}^{-1}NN_r M_c^{-1} = \hat{L}\hat{N}\hat{D}_{cl}^{-1}\hat{N}_r \quad (24)$$

where we introduce a left coprime and a right coprime factorization, respectively, for the transfer matrices $G_c(s)$ and $G(s)$ as follows [11]

$$\begin{aligned} N_c M_c^{-1} &= \bar{M}_c^{-1} \bar{N}_c \quad \text{and} \quad \bar{N}_c N_c + \bar{M}_c M_c = I \\ M^{-1} N &= \bar{N} \bar{M}^{-1} \quad \text{and} \quad \bar{N} \bar{N} + \bar{M} \bar{M} = I \\ \bar{D}_{cl} &= \bar{M}_c \bar{M} + \bar{N}_c \bar{N} \end{aligned}$$

(24) clearly shows that the transfer matrix from w to r also belongs to RH_∞ . For a step-like reference and a failure of the same nature, we deduce

$$\begin{aligned} \lim_{s \rightarrow 0} sr(s) &= [(\hat{L} M_c - \hat{K} N_c) D_{cl}^{-1} N_f]_{s=0} f_0 \\ &+ [\hat{L} M_c D_{cl}^{-1} N N_c M_c^{-1} + \hat{K} N_c D_{cl}^{-1} M]_{s=0} \end{aligned} \quad (25)$$

where $[X]_{s=0}$ indicates that the transfer matrix $X(s)$ is evaluated for $s = 0$. This proves (20) and (21), by identifying Γ and Θ with the corresponding bracketed expression in (25).

Finally, we consider each term of (25) when $\hat{G} \rightarrow G$ and $\hat{G}_f \rightarrow G_f$. We obtain successively

$$\begin{aligned} \lim_{\substack{\hat{G} \rightarrow G \\ \hat{G}_f \rightarrow G_f}} (\hat{L} M_c D_{cl}^{-1} N N_c M_c^{-1} + \hat{K} N_c D_{cl}^{-1} M) \\ = \lim_{\substack{\hat{G} \rightarrow G \\ \hat{G}_f \rightarrow G_f}} Q S_{N_f} \hat{G}_f^{-1} (M_c D_{cl}^{-1} N N_c M_c^{-1} - \hat{G} N_c D_{cl}^{-1} M) \\ = \lim_{\hat{G} \rightarrow G} Q S_{N_f} \hat{G}_f^{-1} (G - \hat{G}) C (I + GC)^{-1} = 0. \end{aligned} \quad (26)$$

$$= \lim_{\hat{G} \rightarrow G} Q S_{N_f} \hat{G}_f^{-1} (G - \hat{G}) C (I + GC)^{-1} = 0. \quad (27)$$

Hence $\Gamma \rightarrow 0$ as $\hat{G} \rightarrow G$. Equation (26) is easily deduced from (12) and (16), while (27) is obtained by straightforward computations.

On the other hand, using (16), we obtain

$$\begin{aligned} \lim_{\substack{\hat{G} \rightarrow G \\ \hat{G}_f \rightarrow G_f}} (\hat{L} M_c - \hat{K} N_c) D_{cl}^{-1} N_f \\ = \lim_{\substack{\hat{G} \rightarrow G \\ \hat{G}_f \rightarrow G_f}} P_{N_f} \hat{M} (I + \hat{G} C) (I + GC)^{-1} G_f \end{aligned} \quad (28)$$

$$= \lim_{\hat{G} \rightarrow G} P_{N_f} \hat{M} G_f + P_{N_f} \hat{M} (\hat{G} - G) C (I + GC)^{-1} G_f \quad (29)$$

$$= \lim_{\hat{G} \rightarrow G} (Q S_{N_f} + Q S_{N_f} \hat{G}_f^{-1} (G_f - \hat{G}_f) + P_{N_f} \hat{M} (\hat{G} - G) C (I + GC)^{-1} G_f) \quad (30)$$

where (30) is deduced from (29) by using (12). The first term in (30) is obviously a diagonal matrix, which concludes the proof. \square

Note that from (19) and (16), (20) and (21) could intuitively be expected. Indeed, as the plant is stabilized, the residual generator needs to be BIBO stable for (20) and (21) to be fulfilled. Since it is designed using the plant model exclusively, model inaccuracies do not affect its stability.

From (27) and (30), we can expect that, when the discrepancy between the estimated and the "true" plant model is sufficiently small, the matrices Γ and Θ will respectively be sufficiently close to zero and to a diagonal matrix to guarantee a satisfactory operation of the residual generator.

V. CONCLUSION

We have analyzed the problem of residual generation for an unstable plant, and we have proved that it cannot be solved for such a plant (as soon as there is a slight modeling uncertainty), unless the

plant is stabilized by an adequate controller. This is due to the fact that the unstable modes of the plant cannot be made unobservable from the residual when the model is not exactly known. As a by-product, we have obtained expressions such as (27) and (30), which could be used to quantify the effect of modeling uncertainties on the performance of a residual generator.

For the sake of simplicity, the results are stated and proved for the extension of the fundamental problem of residual generation, and when the number of failures, q , is equal to the number of measurements, p . In the presence of modeling uncertainties, however, the asymptotic stability of the plant model is also required to be able to solve the fundamental problem of residual generation defined in [8]. As this is the basic problem for the design of a residual generator for any coding set, one deduces that the above stability requirement is also necessary when $q > p$. Yet, in this case one cannot isolate simultaneous failures [8].

ACKNOWLEDGMENT

The authors would like to thank an anonymous reviewer for his constructive comments on a first version of the paper.

REFERENCES

- [1] M. Massoumnia, G. C. Verghese, and A. S. Willsky, "Failure detection and identification," *IEEE Trans. Automat. Control*, vol. 34, no. (3), pp. 316-321, 1989.
- [2] P. Alexandre and M. Kinnaert, "Numerically reliable algorithm for the synthesis of linear fault detection filters based on the geometric approach," in *Proc. 1993 IEEE Conf. Syst., Man, Cybernetics*, 1993, pp. 359-364.
- [3] N. Viswanadham, J. H. Taylor, and E. C. Luce, "A frequency domain approach to failure detection and isolation with application to GE-21 turbine engine control systems," *Contr. Theory Advanced Tech.*, vol. 3, no. 1, pp. 45-72, 1987.
- [4] X. Ding and P. M. Frank, "Fault detection via factorization approach," *Syst. Contr. Lett.*, vol. 14, pp. 433-436, 1990.
- [5] M. Kinnaert and Y. Peng, "Residual generator for sensor and actuator fault detection and isolation," in *Proc. 2nd European Contr. Conf.*, Groningen, 1993, pp. 1970-1974.
- [6] E. Y. Chow and A. S. Willsky, "Analytical redundancy and the design of robust failure detection systems," *IEEE Trans. Automat. Contr.*, vol. AC-29, 7, pp. 603-614, 1984.
- [7] J. Gertler, "Analytical redundancy methods in fault detection and isolation," *Preprints of SAFEPROCESS '91*, pp. 9-21, 1991.
- [8] J. Gertler and R. Monajemy, "Generating directional residuals with dynamic parity equations," *Preprints of the 12th IFAC World Congress*, vol. 7, pp. 505-510, 1993.
- [9] R. J. Patton and J. Chen, "Robust fault detection using eigenstructure assignment: a tutorial consideration and some new results," in *Proc. 30th Conf. Decis. Contr.*, Brighton, 1991, pp. 2242-2247.
- [10] J. E. White and J. L. Speyer, "Detection filter design: Spectral theory and algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-32, no. 7, pp. 593-603, 1987.
- [11] M. Vidyasagar, *Control System Synthesis: A Factorisation Approach*. Cambridge, MA: MIT Press, 1987.

Zeros of Discretized Continuous Systems Expressed in the Euler Operator—An Asymptotic Analysis

Addisu Tesfaye and Masayoshi Tomizuka

Abstract—The structure of the zeros of SISO continuous-time systems, which are discretized via a zero-order hold and expressed in the Euler operator, is studied. In particular, it will be shown that when state-space descriptions of linear SISO continuous-time systems with relative degree ≥ 2 are discretized, then the zero dynamics of the resulting discrete system is singularly perturbed and shows a separation of time scale. The part of the zero dynamics associated with the fast time scale is shown to correspond to the zeros introduced by the sampling process (sampling zeros). An asymptotic formula for this part of the zero dynamics is given, and implications of the result to control design based on pole zero cancellation is discussed.

I. INTRODUCTION

In the design of linear control systems, the existence of unstable zeros makes it difficult to construct some control procedures, such as inverse systems, model matching systems, and model reference adaptive controllers [6]. Furthermore, it restricts the control ability of robust controllers [18]. When we use conventional digital control (i.e., z operator) with zero-order hold (ZOH) input, zeros appear in the discrete-time model even though the continuous-time system is minimal phase [1]. These zeros are commonly referred to in the literature as sampling zeros because they arise from the sampling process. Whether these zeros are stable (minimum phase) or unstable (nonminimum phase) depends mainly on the sampling interval T . Generally at fast sampling rates, the sampling zeros appear outside or on the unit circle while at slower sampling speeds they are stable. This poses a paradox for control engineers since generally a sufficiently small sampling interval T is desired from the control point of view, while the sampling zeros tend to become unstable as T becomes smaller.

Many researchers have sought to address this problem by different techniques. Hagiwara and Araki [3] have proposed a multirate sampling scheme which requires n times sampling between a sample period where n is the order of the plant such that the resulting discrete-time system has no finite zeros. A similar double-delay scheme proposed by Mita *et al.* [10] where control input or output measurements are taken twice in one sampling interval also constructs a discrete-time model with no finite zeros. Such controllers inevitably require intersample control or measurement which makes the design of the control systems comparatively difficult or requires additional hardware. Kabamba [5] has proposed the idea of using generalized sampled data hold functions (GSHF) where the input matrix of the discrete-time model can be chosen freely and hence can avoid the unstable sampling zero problem. While this idea has potential advantages, it seems difficult to generate continuous-time hold functions by a digital computer, although they can be approximated by fast input samplers. Other researchers have sought instead to cope with the unstable sampling zeros by compensating for their effects. Notable among such works is the zero phase error tracking control (ZPETC) proposed by Tomizuka [16]. For constructing inverse systems such as in feedforward applications or model reference control, the essential idea behind ZPETC is direct cancellation of all stable zeros and

compensation for the effects of unstable zeros by phase cancellation over all frequencies while making the overall gain approximately equal to one over a desired frequency range.

Recently, one simple and more direct way of overcoming the unstable sampling zero problem has been developed by Middleton and Goodwin [2], [9]. They propose to use an operator called the delta operator ($\delta = \frac{q-1}{T}$ where q is the shift operator in the time domain). In the frequency domain an equivalent operator called the Euler operator can be similarly defined as $\epsilon = \frac{z-1}{T}$ where z is the conventional z operator for discrete-time models written in difference equation form. One major advantage of using these operators is that the dichotomy between results obtained using continuous and discrete-time control laws are resolved especially in regards to the limiting properties as $T \rightarrow 0$ [9]. Furthermore, the zeros arising from sampling are easily distinguished if these operators are utilized. In [2] these authors recommend that the sampling zeros which are unstable for the T selected be ignored if the sampling interval T is sufficiently small. Although ignoring the zeros introduces an error in the discrete-time model, experimental tests on electromechanical systems [2] and electrohydraulic systems [4] indicate that the resulting adaptive control law give excellent performances. Tesfaye and Tomizuka have also applied this procedure successfully in the model reference tracking control of a single axis NSK direct drive robot arm [14], [15].

In this paper we seek to address the structure of the zeros of discretized linear continuous systems modeled by $\delta(\epsilon)$ operators. In particular it will be shown that, under sampling of a continuous linear single-input/single-output (SISO) time system of relative degree ≥ 2 , the resulting δ model system can be regarded as a regular perturbation of the underlying continuous-time system if the sampling interval is considered as a parameter but that the zero dynamics is singularly perturbed i.e., exhibits two-time scales. The conclusion is that the finite sampling zeros become infinite as $T \rightarrow 0$ while the rest of the zeros tend as a set to the zeros of the underlying continuous-time system. An example depicting the behavior of the zeros for a hypothetical continuous-time system is included to substantiate our results.

II. SISO CONTINUOUS-TIME SYSTEM

We investigate linear SISO systems of the form

$$\begin{aligned}\dot{x} &= A_c x + b_c u \\ y &= c_c^T x \quad x \in \mathbb{R}^n, \quad u, y \in \mathbb{R}\end{aligned}\quad (1)$$

Assume (1) is minimal (both controllable and observable) and has relative degree $\gamma \geq 2$ i.e.,

$$\begin{aligned}c_c^T A_c^{i-1} b_c &= \cdots = c_c^T A_c^{\gamma-1} b_c = 0 \\ c_c^T A_c^{\gamma-1} b_c &\neq 0.\end{aligned}\quad (2)$$

Remark: The case where $\gamma \geq 2$ is considered because continuous-time systems give rise to discrete-time models having nonminimum phase zeros only when the relative degree is greater than one [1]. To characterize the zero dynamics of (1) we transform it into normal form by means of the following change of coordinates

$$[\zeta, \eta]^T = [c_c^T, (c_c A_c)^T, \dots, (c_c A_c^{\gamma-1})^T]^T x \quad (3)$$

where $\zeta \in \mathbb{R}^\gamma$, $\eta \in \mathbb{R}^{n-\gamma}$ and H must be selected so that L_c is nonsingular. For simplicity we will choose H such that $H b_c = 0$. The system description (1) can then be rewritten in terms of the new

Manuscript received September 14, 1993; revised April 1, 1994.

The authors are with the Department of Mechanical Engineering, University of California at Berkeley, Berkeley, CA 94720 USA.

IEEE Log Number 9408798.

coordinates as

$$\begin{aligned}\dot{\zeta}_{i,1} &= \zeta_{i,2}, \dot{\zeta}_{i,2} = \zeta_{i,3}, \dots, \dot{\zeta}_{i,\gamma-1} = \zeta_{i,\gamma} \\ \dot{\zeta}_{i,\gamma} &= R_{\zeta}\zeta_i + S_{\eta}\eta_i + cA_i^{\gamma-1}b_i u \\ \dot{\eta}_i &= P_{\zeta}\zeta_i + Q_{\eta}\eta_i\end{aligned}\quad (4)$$

where $R_{\zeta}^{(i)} \in \mathbb{R}^{\gamma}$, $S_{\eta}^{(i)} \in \mathbb{R}^{n-\gamma}$, $P_{\zeta}^{(i)} \in \mathbb{R}^{(n-\gamma) \times \gamma}$ and $Q_{\eta}^{(i)} \in \mathbb{R}^{(n-\gamma) \times (n-\gamma)}$. $R_{\zeta}\zeta_i + S_{\eta}\eta_i = cA_i^{\gamma}x$, $Q_{\eta}\eta_i + P_{\zeta}\zeta_i = HA_i x$. Equation (4) expresses (1) in normal form, and it is well known that Q_{η} is a companion matrix whose $n - \gamma$ eigenvalues are precisely the zeros of system (1).

III. SAMPLED DATA SYSTEM IN δ FORM

The sampled ZOH representation of system (1) can be written in δ ($\delta = \frac{T}{T}$ where q is the shift operator) form as

$$\begin{aligned}\delta x(k) &= A_{\delta}x(k) + b_{\delta}u(k) \\ y &= c x(k)\end{aligned}\quad (5)$$

where $A_{\delta} = A_c + TA_1$, $b_{\delta} = b_c + Tb_1$

$$\begin{aligned}A_1 &= \frac{1}{2!}A_c^2 + \frac{T}{3!}A_c^3 + \dots + \frac{T^{k-2}}{k!}A_c^{k-1} + \dots \\ b_1 &= \left(\frac{1}{2!}A_c + \frac{T}{3!}A_c^2 + \dots + \frac{T^{k-1}}{(k+1)!}A_c^k + \dots \right) b_c.\end{aligned}$$

Notice that if we regard the sampling time, T , as a perturbation parameter, A_{δ} , b_{δ} can be considered as regular perturbations of A_c and b_c , respectively. Further based on the assumption of minimality of (1) we conclude that (5) remains minimal for T small. Consider a transformation of (5) to normal coordinates

$$\begin{aligned}[\zeta^T \eta^T]^T &= \\ &\underbrace{\left[c^T, (c(A_c + TA_1))^T, \dots, (c(A_c + TA_1)^{\gamma-1})^T, H^T \right]^T}_{L} x.\end{aligned}\quad (6)$$

Again notice that matrix L of the above equation is a regular perturbation of matrix L_c of (3) and that $\det(L) \neq 0$ for T sufficiently small. Let

$$A_1 = A_c A_1^*, \quad b_1 = b_1^* b_c$$

where

$$\begin{aligned}A_1^* &= \frac{1}{2!}A_c + \frac{T}{3!}A_c^2 + \dots + \frac{T^{k-2}}{k!}A_c^{k-1} + \dots \\ b_1^* &= \frac{1}{2!}A_c + \frac{T}{3!}A_c^2 + \dots + \frac{T^{k-2}}{k!}A_c^{k-1} + \dots\end{aligned}$$

i.e.,

$$\begin{aligned}A_1^* &= b_1^* \\ A_c + TA_1 &= A_c (I + Tb_1^*) \\ b_c + Tb_1 &= (I + Tb_1^*) b_c.\end{aligned}$$

In the subsequent analysis we will repeatedly use the properties given by (2) of the underlying continuous system (1). Considering the transformation of b_{δ} we have for each row of Lb_{δ}

$$\begin{aligned}0) \quad c(b_c + Tb_1) &= T^{\gamma-1}\alpha_0 + T^2\{\cdot\}\end{aligned}$$

where

$$\alpha_0 = \frac{1}{\gamma!}cA_c^{\gamma-1}b_c.$$

1)

$$\begin{aligned}c(A_c + TA_1)(b_c + Tb_1) &= cA_c(I + Tb_1^*)^2 b_c \\ &= T^{\gamma-1}\alpha_{11} + T^{\gamma-2}\alpha_{12} + T^2\{\cdot\}\end{aligned}$$

where

$$\alpha_{11} = \frac{2}{\gamma!}cA_c^{\gamma}b_c, \quad \alpha_{12} = \frac{2}{(\gamma-1)!}cA_c^{\gamma-1}b_c.$$

2)

$$\begin{aligned}c(A_c + TA_1)^2(b_c + Tb_1) &= cA_c^2(I + Tb_1^*)^3 b_c \\ &= T^{\gamma-2}\alpha_{21} + T^{\gamma-1}\alpha_{22} + T^2\{\cdot\}\end{aligned}$$

where

$$\alpha_{21} = \frac{3}{(\gamma-1)!}cA_c^{\gamma}b_c, \quad \alpha_{22} = \frac{3}{(\gamma-2)!}cA_c^{\gamma-1}b_c.$$

⋮

• i)

$$\begin{aligned}c(A_c + TA_1)^i(b_c + Tb_1) &= cA_c^i(I + Tb_1^*)^{i+1} b_c \\ &= T^{\gamma-i}\alpha_{i1} + T^{\gamma-i+1}\alpha_{i2} + T^2\{\cdot\}\end{aligned}$$

where

$$\alpha_{i1} = \frac{(i+1)}{(\gamma-i+1)!}cA_c^{\gamma}b_c, \quad \alpha_{i2} = \frac{(i+1)}{(\gamma-i)!}cA_c^{\gamma-1}b_c.$$

• (i-1)

$$\begin{aligned}c(A_c + TA_1)^{\gamma-1}(b_c + Tb_1) &= cA_c^{\gamma-1}\left(b_c + \frac{\gamma T}{2}A_1 b_c\right) \\ &\quad + O(T^2)\end{aligned}$$

Also

$$H(b_c + Tb_1) = \frac{T}{2}HA_c + O(T^2)$$

The expression $T^2\{\cdot\}$ denotes some function which we drop because it will not affect the asymptotic solutions to be developed. Even after dropping the expression $T^2\{\cdot\}$ we note, from the above, that the first $\gamma - 2$ rows of Lb_{δ} have coefficients which are at least of order $O(T^1)$. At this stage, in view of facilitating an asymptotic analysis which: 1) captures the overall behavior without sacrificing the structure of Lb_{δ} , and 2) avoids a cumbersome analysis involving a multiplicity of time scales, we replace the first $\gamma - 2$ rows of Lb_{δ} by bounding functions of the form $T\alpha'_0, T\alpha'_1, \dots, T\alpha'_{\gamma-3}$ where the α'_i 's are independent of T . For example, a particular choice for the α'_i 's is $\alpha'_0 = \alpha_0, \alpha'_1 = \alpha_{11} + \alpha_{12}, \alpha'_{\gamma-3} = \alpha_{(\gamma-3)1} + \alpha_{(\gamma-3)2}$. The above simplification will not affect the final conclusions and is justified because we seek asymptotic solutions. Taking this into consideration, a simplified expression which captures the behavior of Lb_{δ} (with respect to T) can be expressed as

$$\begin{bmatrix} T\alpha'_0 \\ T\alpha'_1 \\ \vdots \\ T\alpha'_{\gamma-3} \\ T^{\frac{\gamma-1}{2}}cA_c^{\gamma-1}b_c \\ cA_c^{\gamma-1}(b_c + \frac{\gamma T}{2}A_1 b_c) \\ \frac{T}{2}HA_c \end{bmatrix} \quad (7)$$

Note that if $T = 0$, the above matrix equals $L_c b_c$, the transformed continuous-time input matrix (i.e., applying (2) and (3)). It is obvious from this last expression that if we transform system (5) into normal

form by means of the transformation given by (6) the feedback law

$$u = \frac{-1}{T\alpha'_0} \zeta_2 = \frac{-1}{T\alpha'_0} c(A_c + T A_1)x \quad (8)$$

makes

$$y(k) = \delta y(k) = 0$$

and that the zero dynamics subspace is

$$\begin{aligned} \nu &= \{x: y = cx = 0\} \\ &= \{(0, \bar{\zeta}, \eta) : \bar{\zeta}^T = [\bar{\zeta}_1 \ \bar{\zeta}_2 \ \cdots \ \bar{\zeta}_{\gamma-1}] \\ &\quad \in \mathbb{R}^{\gamma-1}, \eta \in \mathbb{R}^{n-\gamma}\}. \end{aligned}$$

It should be noted that ζ has been decomposed as $\zeta^T = [\zeta_1 \ \bar{\zeta}^T]$ i.e., in the last expression $\bar{\zeta}_1 = c(A_c + T A_1)x$, $\zeta_2 = c(A_c + T A_1)^2 x$, \dots , $\bar{\zeta}_{\gamma-1} = c(A_c + T A_1)^{\gamma-1} x$. The most important thing to observe from (7) is that all the entries of the matrix are multiplied by the sampling time T except for the one next to the last row block. On the subspace ν , the dynamics, under the feedback law (8), is given by

$$\begin{aligned} \delta \zeta(k) &= \begin{bmatrix} \frac{-\alpha'_1}{\alpha'_0} & 1 & 0 \\ \frac{\alpha'_2}{\alpha'_0} & 0 & 1 & 0 \\ \vdots & & & \\ \frac{\alpha'_{\gamma-1}}{\alpha'_0} & 0 & & 1 \\ -\frac{(c A_c^{\gamma-1} b_c)}{2\alpha'_0} & & & & 1 \\ \frac{(c A_c^{\gamma-1} b_c + \frac{\gamma-1}{2} \alpha'_1 b_c)}{T\alpha'_0} + r_2 & r_1 & \cdots & \cdots & r_{\gamma-1} \end{bmatrix} \bar{\zeta}(k) \\ &+ \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ s \end{bmatrix} \eta(k) \quad (9) \end{aligned}$$

$$\delta \eta(k) = \left[\left(p_2 - \frac{H A_c}{2\alpha'_0} \right) p_1 \ \cdots \ p_{\gamma-1} \right] \bar{\zeta}(k) + Q_\eta \eta(k). \quad (10)$$

In the above $r_2, r_1, \dots, r_{\gamma-1}, s, p_2, p_1, \dots, p_{\gamma-1} \in \mathbb{R}$. From the above expression, for a sampling time $T \neq 0$, we observe that the ZOH equivalent representations of continuous-time systems expressed in the Euler (ϵ) operator will have relative degree one regardless of the relative degree of the underlying continuous-time system. This means that such representations always lead to $n-1$ zeros. These zeros are composed of: 1) $n-\gamma$ zeros which are due to the continuous-time system, and 2) $\gamma-1$ zeros which are a consequence of the sampling

process. The latter zeros are commonly referred to as sampling zeros. To establish the structure of the sampling zeros we note from (9) that the T dependent term corresponding to $\delta \bar{\zeta}_{\gamma-1}(k)$ (i.e., last row of $\delta \bar{\zeta}(k)$) is of the order of $1/T$ and thus is not a regular perturbation but rather a singular perturbation. This singular perturbation is due to the high-gain form of the feedback control (8) (division by T). High-gain systems arising from small regular perturbations have been investigated by many researchers in particular in [12] and [17] in continuous-time systems and by Rao and Naidu in discrete-time systems [11]. More recently Sastry and Hauser [13] have investigated the appearance of such singular perturbations arising from regular perturbations in both linear and nonlinear systems. The following conclusion is analogous to Theorem 2.1 of [13].

Theorem 1 Consider a continuous-time system of the form (1) and its ZOH equivalent representation given by (5) which is expressed by means of the δ operator. Then if system (1) has relative degree γ , system (5) has $n-1$ zeros which according to their asymptotic behavior as $T \rightarrow 0$ belong to two groups:

- 1) The $\gamma-1$ sampling zeros tend to ∞ asymptotically as

$$\left(\frac{-1}{T} \frac{c A_c^{\gamma-1} b_c}{\alpha'_0} \right)^{\frac{1}{\gamma-1}}.$$

- 2) The remaining $n-\gamma$ zeros tend to zeros of the continuous-time system given by (1).

Proof. The proof is straightforward if we can cast (9) and (10) into a standard singular perturbation form. This can be accomplished by rescaling $\bar{\zeta}$ in the following manner. Denote

$$\begin{aligned} z_1 &= \bar{\zeta}_1, \quad z_2 = T^{\frac{1}{\gamma-1}} \bar{\zeta}_2, \quad z_3 = T^{\frac{2}{\gamma-1}} \bar{\zeta}_3 \\ z_i &= T^{\frac{i-1}{\gamma-1}} \bar{\zeta}_i, \quad \dots, \quad z_{\gamma-1} = T^{\frac{\gamma-2}{\gamma-1}} \bar{\zeta}_{\gamma-1}. \end{aligned}$$

This expresses a transformation (rescaling) of (9) and (10) which can be written in matrix form as

$$\begin{bmatrix} \bar{\zeta} \\ \eta \end{bmatrix} = M \begin{bmatrix} z \\ \eta \end{bmatrix} \quad (11)$$

where

$$\begin{aligned} M &= \begin{bmatrix} M_1 & M_2 \\ M_3 & M_4 \end{bmatrix} \\ M_1 &= \begin{bmatrix} 1 & 0 \\ 0 & T^{\frac{1}{\gamma-1}} & & \\ & T^{\frac{2}{\gamma-1}} & 0 & \\ & 0 & \ddots & \\ & & & T^{\frac{\gamma-2}{\gamma-1}} \end{bmatrix} \\ M_2 &= 0_{(\gamma-1) \times (n-\gamma)}, \quad M_3 = 0_{(n-\gamma) \times (\gamma-1)} \\ M_4 &= I_{(n-\gamma) \times (n-\gamma)}. \end{aligned}$$

$$\begin{aligned} T^{\frac{1}{\gamma-1}} \delta z_1(k) &= -T^{\frac{1}{\gamma-1}} \frac{\alpha'_1}{\alpha'_0} z_1(k) + z_2(k) \\ T^{\frac{1}{\gamma-1}} \delta z_2(k) &= z_1(k) + O(T^{\frac{2}{\gamma-1}}) \\ &\vdots \\ T^{\frac{1}{\gamma-1}} \delta z_{\gamma-2}(k) &= z_{\gamma-1}(k) + O(T^{\frac{\gamma-2}{\gamma-1}}) \\ T^{\frac{1}{\gamma-1}} \delta z_{\gamma-1}(k) &= \frac{-c A_c^{\gamma-1} b_c}{\alpha'_0} z_1(k) + T^{\frac{1}{\gamma-1}} r_\gamma z_{\gamma-1}(k) + \\ &\quad -T \left(\frac{\gamma T}{2} \frac{c A_c^{\gamma-1} b_c}{\alpha'_0} + r_2 \right) z_1(k) + T^{\frac{\gamma-2}{\gamma-1}} r_3 z_2(k) + \cdots + T^{\frac{\gamma-1}{\gamma-1}} r_{\gamma-1} z_{\gamma-2}(k) + T s \eta(k) \\ &\quad \underbrace{\hspace{10em}}_{O(T^{\frac{\gamma-1}{\gamma-1}})} \end{aligned}$$

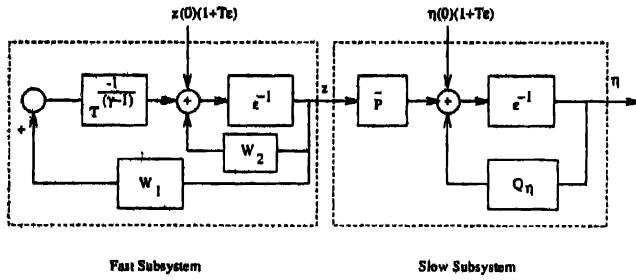


Fig. 1. Singularly perturbed (two-time scale) zero dynamics.

For simplicity let us express (9) and (10) as

$$\begin{aligned}\delta\tilde{\zeta}(k) &= A_1\tilde{\zeta}(k) + A_2\eta(k) \\ \delta\eta(k) &= A_3\tilde{\zeta}(k) + A_4\eta(k).\end{aligned}\quad (12)$$

Combining (11) and (12) we can write (9) and (10) in terms of the new state variables as

$$\begin{bmatrix} \delta z(k) \\ \vdots \\ \delta\eta(k) \end{bmatrix} = \begin{bmatrix} M_1 A_1 M_1^{-1} & M_1 A_2 \\ A_3 M_1^{-1} & A_4 \end{bmatrix} \begin{bmatrix} z(k) \\ \vdots \\ \eta(k) \end{bmatrix}\quad (13)$$

$$M_1 A M_1^{-1} = \begin{bmatrix} \frac{-\alpha'_1}{\alpha'_0} & T^{\frac{-1}{\gamma-1}} & 0 \\ -T^{\frac{-1}{\gamma-1}} \frac{\alpha'_2}{\alpha'_0} & 0 & T^{\frac{-1}{\gamma-1}} & 0 \\ -T^{\frac{-2}{\gamma-1}} \frac{\alpha'_3}{\alpha'_0} & 0 & 0 & \ddots \\ \vdots & & \vdots & T^{\frac{-1}{\gamma-1}} \\ -T^{\frac{-\gamma}{\gamma-1}} \{\bar{r}_2 + r_2\} & T^{\frac{-1}{\gamma-1}} r_3 & T^{\frac{-4}{\gamma-1}} r_4 & \cdots & r_\gamma \end{bmatrix} \tilde{\zeta}(k)$$

$$\bar{r}_2 = \frac{c A_i^{\gamma-1} (b_i + \frac{\gamma}{2} A_i b_i)}{T \alpha'_0}.$$

Equivalently we have the equation found at the bottom of the previous page. Note that $M_1 A_2 = T^{\frac{\gamma-1}{\gamma-1}} \bar{s}$ and (13) can be concisely written as

$$\begin{bmatrix} T^{\frac{-1}{\gamma-1}} \delta z(k) \\ \delta\eta(k) \end{bmatrix} = \begin{bmatrix} W_1 & 0 \\ \tilde{P} & Q \end{bmatrix} \begin{bmatrix} z(k) \\ \eta(k) \end{bmatrix} + \begin{bmatrix} T^{\frac{-1}{\gamma-1}} W_2 z(k) \\ 0 \end{bmatrix} + \begin{bmatrix} O(T^{\frac{-4}{\gamma-1}}) \\ 0 \end{bmatrix}\quad (14)$$

$$\tilde{P} = A_1 M_1^{-1}$$

where

$$W_1 = \begin{bmatrix} 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ & & & \ddots & \\ & & & 0 & 1 \\ \frac{-c A_i^{\gamma-1} b_i}{\alpha'_0} & 0 & \cdots & 0 & 0 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} \frac{-\alpha'_1}{\alpha'_0} & 0 & & & \\ 0 & 0 & & & \\ & & & \ddots & \\ & & 0 & & 0 \\ 0 & \cdots & 0 & & r_\gamma \end{bmatrix}.$$

From the above, (14) is observed to be in the standard two-time scale form of [7]. Its right-hand side is regularly perturbed by the term $T^{\frac{-1}{\gamma-1}} W_2$, while the matrix of the unperturbed part is block lower triangular. This unperturbed matrix is nonsingular as required for a standard two-time scale form. It follows that the eigenvalues of (9) are asymptotically

$$(T^{\frac{-1}{\gamma-1}} \lambda(W_1)) \cup \lambda(Q_\eta)$$

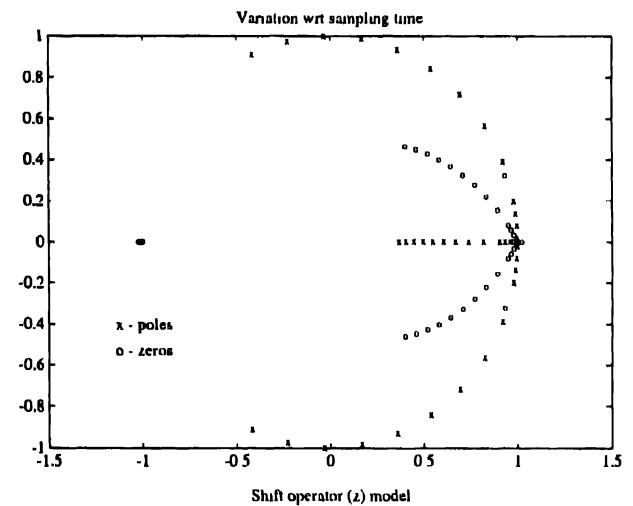
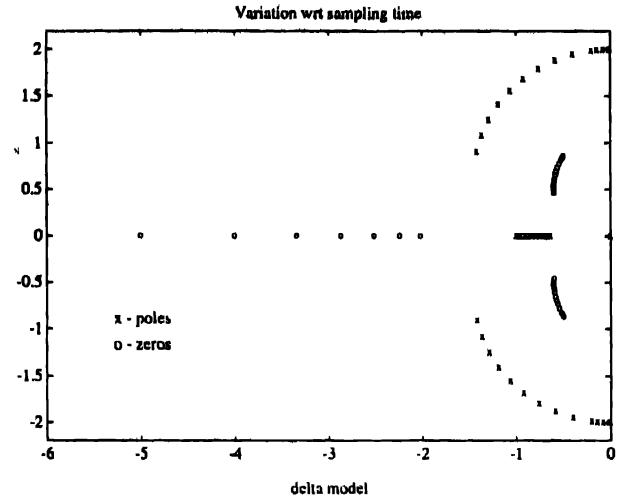


Fig. 2. Pole/zero variation with respect to sampling time

$$(T^{\frac{-1}{\gamma-1}} \lambda(W_1)) = \left(\frac{-1}{T} \frac{c A_i^{\gamma-1} b_i}{\alpha'_0} \right)^{\frac{-1}{\gamma-1}} \quad (15)$$

where (15) is the asymptotic expression of the $\gamma - 1$ sampling zeros which tend to ∞ as $T \rightarrow 0$. This completes the proof of part 1) of the theorem. To prove part 2) observe that from (14) the remaining $n - \gamma$ zeros tend as a set to the eigenvalues of (Q_η) which are the zeros of the underlying system. Following the standard two-time scale form of [7] we can depict the behavior of the zero dynamics of system (5) as composed of a fast and a slow subsystem as follows (the external inputs characterize the effect of initial conditions).

Remark: Theorem 1 states that $\gamma - 1$ sampling zeros become infinite as $T \rightarrow 0$. Moreover one can deduce that the sampling zeros move in specific directions as determined by (15). The consequence of Theorem 1 is that sampling zeros of discrete-time models expressed in the $\delta(\epsilon)$ operator are located far off in the right- or left-half planes for sufficiently small sampling times and hence are slightly nonminimum phase. This supports the recommendation made by Goodwin *et al.* [2] to discard the unstable sampling zeros if T is sufficiently small.

IV. ILLUSTRATIVE EXAMPLE

The following example is taken from [8]. Consider the following transfer function of a continuous-time system

$$G(s) = \frac{s^2 + s + 1}{s(s+1)(s^2+4)}.$$

The system has relative degree two with zeros located at $\{-0.5 \pm 0.866j\}$ while the poles are at $\{0, \pm 2j, -1\}$. Fig. 1 portrays the influence of the sampling time, T , on the location of the poles and zeros for the delta and shift operator (z) models. We notice for the delta model that the sampling zero departs to $-\infty$ as $T \rightarrow 0$ while the rest of the poles and zeros converge to the continuous-time poles and zeros, respectively. This behavior supports Theorem 1 above which predicts that the sampling zero goes to $-\frac{2}{T}$ since in this case $\alpha'_0 = \alpha_0 = \frac{1}{\gamma} cA^{-1}b$. For the shift operator model we notice that the zero introduced by sampling remains fixed around -1 while the rest of the poles and zeros converge to one on the real axis irrespective of the locations of the poles and zeros of the underlying continuous-time system. To give an indication of the behavior of the sampling zero in the delta model we calculate for $T = 0.1, 0.07, 0.04, 0.01$, and 0.001 secs. the location of the zero at $-20, -28.6, -50, -200$, and -2000 , respectively. The numerical data validates Theorem 1 since the sampling zero goes to $-\frac{2}{T}$.

V. CONCLUSION

The effect of the sampling time on the zero dynamics of continuous linear SISO systems expressed in the δ operator has been investigated. It is shown that as the sampling time tends to zero that the $\gamma - 1$ zeros introduced by sampling (sampling zeros) migrate to negative infinity while the remaining $n - \gamma$ zeros converge uniformly to that of the underlying continuous-time system. From this we can conclude that under fast sampling, i.e., as $T \rightarrow 0$, one can neglect the sampling zeros (which are easily distinguished). This conclusion is important for control systems constructed based on system inversion and/or model reference control systems especially when the dynamics of the plant are imperfectly known.

REFERENCES

- [1] K. J. Astrom, P. Hagander, and J. Sienby, "Zeros of sampled systems," *Automatica*, vol. 20, pp. 31-38, 1984.
- [2] G. C. Goodwin, R. Lozano-Leal, D. Q. Mayne, and R. H. Middleton, "Rapprochement between continuous and discrete model reference adaptive control," *Automatica*, vol. 22, pp. 199-207, 1986.
- [3] T. Hagiwara and M. Araki, "Design of a stable state feedback controller based on the multirate sampling of the plant output," *IEEE Trans Automat Contr.*, vol. 33, pp. 812-819, Oct. 1988.
- [4] N. Horn, D. V. Bitner, P. N. Nikiforuk and P. R. Ukrainetz, "Robust adaptive control of electrohydraulic servo systems using the Euler operator," in *Proc. Int Conf Contr.*, vol. 1, Edinburgh, 1991, pp. 671-676.
- [5] P. T. Kabamba, "Control of linear systems using generalized sampled-data hold functions," *IEEE Trans Automat Contr.*, vol. AC-32, pp. 772-783, 1987.
- [6] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1980.
- [7] P. Kokotovic, H. Khalil, and J. O'Reilly, *Singular Perturbations in Control Analysis And Design*. New York: Academic, 1986.
- [8] R. H. Middleton and G. C. Goodwin, *Digital Control And Estimation—A Unified Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1990.
- [9] —, "Improved finite word length characteristics in digital control using Delta operators," *IEEE Trans. Automat. Contr.*, vol. 31, no. 11, pp. 1015-1021, 1986.
- [10] T. Mita, Y. Chida, Y. Kaku, and H. Numasato, "Two-delay robust digital control and its applications—avoiding the problem on unstable limiting zeros," *IEEE Trans Automat Contr.*, vol. 35, no. 8, pp. 962-970, 1990.
- [11] A. K. Rao and D. S. Naidu, "Singularly perturbed boundary value problems in discrete time systems," *Int J. Contr.*, vol. 34, 1981.
- [12] P. Sannuti, "Direct singular perturbations analysis of high-gain and cheap control problems," *Automatica*, vol. 19, pp. 41-51, 1983.
- [13] S. Sastry, J. Hauser, and P. Kokotovic, "Zero dynamics of regularly perturbed systems may be singularly perturbed," *Syst. Contr. Lett.*, vol. 13, pp. 299-314, 1989.
- [14] A. Tesfaye and M. Tomizuka, "Robust digital tracking with perturbation estimation via the Euler operator," in *Proc. Winter Ann. Meet. ASME, DSC*, vol. 50, 1993, pp. 49-54.
- [15] —, "Sliding control of discretized continuous systems via the Euler operator," in *Proc. 32nd IEEE Conf Decis Contr.*, vol. 1, 1993, pp. 871-876.
- [16] M. Tomizuka, "Zero-phase error tracking controller for digital control," *ASME J. Dynamic Sys., Measurement, Contr.*, vol. 109, pp. 65-68, 1987a.
- [17] K. Young, P. Kokotovic, and V. Utkin, "A singular perturbation analysis of high-gain feedback systems," *IEEE Trans. Automat. Contr.*, vol. 22, pp. 931-938, 1977.
- [18] M. Vidyasagar, *Control System Synthesis—A Factorization Approach*. Cambridge, MA: MIT Press, 1985.

Further Theoretical Results on Direct Strain Feedback Control of Flexible Robot Arms

Zheng-Hua Luo and Baozhu Guo

Abstract—This paper is concerned with stability analyses for some nonstandard second-order partial differential equations arising from direct strain feedback control of flexible robot arms. Exponential stability issues are addressed for three types of differential equations, one of which is in general abstract evolution equation form and the other two are in partial differential equation form. The obtained results are of especially theoretical interest because they reveal the essence of direct strain feedback control and demonstrate its power in control of flexible arms.

I. INTRODUCTION

We consider the following initial-boundary problem

$$\begin{aligned} \ddot{w}(t, x) + u''''(t, x) + kx\dot{w}(t, 0) &= 0 \\ w(t, 0) &= w'(t, 0) = 0 \\ u''(t, t) &= u'''(t, t) = 0 \end{aligned} \quad (1)$$

$$w(0, x) = w_0(x), \quad \dot{w}(0, x) = w_1(x), \quad x \in (0, l).$$

Equation (1) represents a closed-loop dynamic equation of a single-link flexible arm with the so-called direct strain feedback (DSFB) control [9], [10]. The symbol $w(t, x)$ denotes the arm's bending vibration deflection at time t and position x ; \ddot{w} and u'''' represent time and spacial derivatives of w , respectively. The symbols l and $k > 0$ are, respectively, the length of the arm and strain feedback gain. $w_0(x)$ and $w_1(x)$ are respectively initial position and initial velocity of flexible arms. Originally, the dynamic model of a single-link flexible arm rotating in the horizontal plane by a control motor is given by the following partial differential equation

$$\ddot{w}(t, x) + u''''(t, x) = -x\ddot{\theta}(t)$$

together with the same initial and boundary conditions as in (1), where $\ddot{\theta}(t)$ represents motor rotating acceleration [16]. Note that here a nonlinear term $x\ddot{\theta}(t)^2 w(t, x)$ which should appear on the right-hand side of the above equation, is neglected to obtain a linear model, as in most literature [1] and [16]. The simple yet efficient DSFB control

Manuscript received March 14, 1994.

Z.-H. Luo is with the Department of Mechanical Engineering, Nagaoka University of Technology, Nagaoka, Niigata 940-21, Japan.

B. Guo is with the Department of Applied Mathematics, Beijing Institute of Technology, Beijing 100081, China.

IEEE Log Number 9408799.

is to control the motion of the motor in such a way that $\ddot{\theta}(t)$ is proportional to $\dot{w}''(t, 0)$. This objective is easily achieved by direct feedback of the bending moment $w''(t, 0)$ which can be measured by cementing strain gauge foils at the root end of the arm. The interested reader is referred to [9] and [10] for details.

We are also interested in an equation

$$\begin{cases} \ddot{w}(t, x) + w''''(t, x) + kx\dot{w}''(t, 0) = x(k_1\dot{e}(t) + k_2e(t)) \\ \dot{e}(t) + k_1\dot{e}(t) + k_2e(t) = k\dot{w}''(t, 0) \\ w(t, 0) = w'(t, 0) = w''(t, \ell) = w'''(t, \ell) = 0 \\ w(0, x) = w_0(x), \quad \dot{w}(0, x) = w_1(x) \end{cases} \quad (2)$$

which arises from simultaneous vibration/motion control of flexible arms. More precisely, if we not only control vibration of the arm, but also control the motor position $\theta(t)$, then we can command the motor to move so that $\dot{\theta}(t) = -k_1\dot{e}(t) - k_2e(t) + k\dot{w}''(t, 0)$, where $e(t) (= \theta(t) - \theta_d)$ denotes difference between current motor position $\theta(t)$ and a set-point position θ_d , and $k_1 > 0, k_2 > 0$ are servo feedback gain constants.

Let $H = L^2(0, \ell)$ be the usual square integrable function space, with inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\| \cdot \|$. For later use, let us first introduce an operator A on H as

$$\begin{cases} D(A) = \{w \mid w \in H, w'''' \in H, \\ w(0) = w'(0) = w''(\ell) = w'''(\ell) = 0\} \\ Aw = w'''' \quad \forall w \in D(A) \end{cases}$$

which is known to be positive definite with a compact inverse. Also let us define an operator Π by

$$\Pi w = xw''(0), \quad x \in (0, \ell), \forall w \in D(\Pi).$$

It is obvious that operator Π is neither self-adjoint nor positive definite. It is shown, however, that Π is A -symmetric and A -positive semi-definite [9], [10]. So that Π can be decomposed as $\Pi = Q\dot{A}$ when restricted on $D(A)$, with Q being a bounded, positive semi-definite operator. By making use of these properties of the above mentioned operators, (1) can be written as an abstract equation on H as follows

$$\begin{cases} \ddot{w}(t) + kQ\dot{A}\dot{w}(t) + Aw(t) = 0 \\ w(0) = w_0, \quad \dot{w}(0) = w_1 \end{cases} \quad (3)$$

We say that (1)–(3) are nonstandard since they possess a special damping term $x\dot{w}''(t, 0)$, or equivalently $Q\dot{A}\dot{w}(t)$, which does not appear elsewhere in the existing literature, although some special cases where $Q = I, A^{-1/2}$, or A^{-1} have appeared in [3], [5], and [7].

Let $H_1 = D(A^{1/2}) \times H$ and $H_2 = D(A) \times D(A^{1/2})$. Then H_1 and H_2 are Hilbert spaces when equipped with the following inner products and the corresponding induced norms

$$\begin{aligned} \left\langle \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}, \begin{bmatrix} \tilde{h}_1 \\ \tilde{h}_2 \end{bmatrix} \right\rangle_{H_1} &= \langle A^{1/2}h_1, A^{1/2}\tilde{h}_1 \rangle + \langle h_2, \tilde{h}_2 \rangle \\ \left\langle \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}, \begin{bmatrix} \tilde{h}_1 \\ \tilde{h}_2 \end{bmatrix} \right\rangle_{H_2} &= \langle Ah_1, A\tilde{h}_1 \rangle + \langle A^{1/2}h_2, A^{1/2}\tilde{h}_2 \rangle. \end{aligned}$$

Now let $z(t) = \begin{bmatrix} A^{1/2}w(t) \\ A^{1/2}\dot{w}(t) \end{bmatrix}$. Then (3) can be rewritten as

$$\dot{z}(t) = \tilde{A}z(t)$$

$$z(0) = \begin{bmatrix} A^{1/2}w_0 \\ A^{1/2}w_1 \end{bmatrix}$$

on H_1 , where operator \tilde{A} is defined by

$$\begin{aligned} D(\tilde{A}) = \left\{ h = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \mid h_2 \in D(A^{1/2}), A^{1/2}h_1 \right. \\ \left. + kQA^{1/2}h_2 \in D(A^{1/2}) \right\} \end{aligned}$$

$$\tilde{A}h = \begin{bmatrix} I & 0 \\ 0 & -A^{1/2} \end{bmatrix} \begin{bmatrix} 0 & I \\ A^{1/2} & kQA^{1/2} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix},$$

$$\forall h \in D(\tilde{A}).$$

Such defined $D(\tilde{A})$ is dense in H_1 , and \tilde{A} is a closed operator. Thus the adjoint operator \tilde{A}^* of \tilde{A} can be uniquely defined. Moreover, it can be shown that, for any $h \in D(\tilde{A})$ and $h^* \in D(\tilde{A}^*)$, there hold $\langle \tilde{A}h, h^* \rangle_{H_1} \leq 0$ and $\langle \tilde{A}^*h^*, h \rangle_{H_1} \leq 0$. Consequently, there exists a strongly continuous semigroup of contractions $S(t)$ generated by \tilde{A} [4], [14], and for initial conditions w_0 and w_1 satisfying $w_0, w_1 \in D(A)$ and $Aw_0 + kQA^{1/2}w_1 \in D(A^{1/2})$, there exists a unique solution to (3) which can be expressed as

$$\begin{bmatrix} w(t) \\ \dot{w}(t) \end{bmatrix} = T(t) \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad (4)$$

where

$$T(t) = \begin{bmatrix} A^{-1/2} & 0 \\ 0 & A^{-1/2} \end{bmatrix} S(t) \begin{bmatrix} A^{1/2} & 0 \\ 0 & A^{1/2} \end{bmatrix} \quad (5)$$

is a strongly continuous semigroup on H_2 (this is easily verified). Also, it can be shown that the solution is asymptotically stable, if the operator Q is assumed to be positive definite.

It is the purpose of this paper to present some newly obtained stability results for (1)–(3). Specifically, in Section II, we want to show exponential stability of the solution of (3), under the positivity assumption of Q . This assumption can always be satisfied in flexible arm control if we consider the effect of the Kelvin-Voigt damping in arm's dynamics. The motivation of this section is to see whether the exponential stability result can be extended to the (3) which represents a wide class of systems, not just the DSFB controlled systems. For the specific (1), what is theoretically more interesting is whether we can show exponential stability without introducing any natural damping (internal damping in the material). This question is completely answered in Section III where we show exponential stability of (1). The significance of the obtained results is that they reveal the essence of direct feedback control and demonstrate its power in control of flexible arms. Finally, we provide a sufficient condition on feedback gain k_2 to guarantee the stability of (2), since this is a hybrid system in the sense that its state space consists of an infinite dimensional part and a finite dimensional part, and it is very hard to find a Lyapunov function for the entire system.

II. EXPONENTIAL STABILITY OF (3)

In this section, we study the exponential stability of (3) under the assumption that the operator Q is positive definite. This assumption is always satisfied in flexible arm control, if we take into account the effect of the Kelvin-Voigt viscous natural damping in arm's dynamics. An important reason that we want to do analyses here is that (3) may be considered as an abstract equation of a wide class of systems, not just flexible arm systems as described in [10], and there does not exist exponential results for such general abstract equations, although some results have been obtained for some specific boundary controlled partial differential equations [2], [12]. The energy multiplier method proposed by Chen [2] will be adopted.

Let

$$E(t) = \frac{1}{2} \begin{bmatrix} w(t) \\ \dot{w}(t) \end{bmatrix}_{H_2}^2 = \frac{1}{2} \|Aw(t)\|^2 + \frac{1}{2} \|A^{1/2}\dot{w}(t)\|^2 \quad (6)$$

be an energy function for (3). Then its time derivative along solutions of (3) is given by

$$\dot{E}(t) = -k\langle Q\dot{A}\dot{w}(t), A\dot{w}(t) \rangle \leq 0 \quad (7)$$

which means that the energy will be dissipating. Choose a positive constant $0 < \varepsilon < 1$ and define

$$V(t) = 2(1 - \varepsilon)tE(t) + \langle \dot{w}(t), Aw(t) \rangle. \quad (8)$$

Since

$$\langle \dot{w}(t), Aw(t) \rangle \leq \frac{1}{2} (\|A^{-1/2}\|^2 \|A^{1/2} \dot{w}(t)\|^2 + \|Aw(t)\|^2)$$

always holds, there exists a constant C such that

$$(2(1-\varepsilon)t - C)E(t) \leq V(t) \leq (2(1-\varepsilon)t + C)E(t). \quad (9)$$

Obviously $V(t)$ is a positive function for $t > T_1 := C/2(1-\varepsilon)$.

Considering now the time derivative of $V(t)$ along solutions of (3), we obtain

$$\begin{aligned} \dot{V}(t) &= 2(1-\varepsilon)E(t) + 2(1-\varepsilon)t\dot{E}(t) \\ &\quad + \langle \ddot{w}(t), Aw(t) \rangle + \langle \dot{w}(t), A\dot{w}(t) \rangle \\ &= (1-\varepsilon)\|A^{1/2}\dot{w}(t)\|^2 + (1-\varepsilon)\|Aw(t)\|^2 \\ &\quad - 2k(1-\varepsilon)t\langle Q\dot{w}(t), A\dot{w}(t) \rangle \\ &\quad - k\langle Q\dot{w}(t), Aw(t) \rangle - \langle Aw(t), Aw(t) \rangle \\ &\quad + \langle \dot{w}(t), A\dot{w}(t) \rangle \\ &= (2-\varepsilon)\|A^{1/2}\dot{w}(t)\|^2 - \|Aw(t)\|^2 \\ &\quad - 2k(1-\varepsilon)t\langle Q\dot{w}(t), A\dot{w}(t) \rangle \\ &\quad - k\langle Q\dot{w}(t), Aw(t) \rangle. \end{aligned}$$

Denote now $\lambda_{\max}(Q) = \max_{w \in H} \langle Qw, w \rangle$ and $\lambda_{\min}(Q) = \min_{w \in H} \langle Qw, w \rangle$. Then for an arbitrary constant $a \neq 0$, it holds

$$\begin{aligned} -\langle Q\dot{w}(t), Aw(t) \rangle &\leq \lambda_{\max}(Q)\|\dot{w}(t)\|\|Aw(t)\| \\ &\leq \frac{1}{2}\lambda_{\max}(Q)\left(\|\dot{w}(t)\|^2 + \frac{1}{a^2}\|Aw(t)\|^2\right). \end{aligned}$$

Consequently

$$\begin{aligned} \dot{V}(t) &\leq (2-\varepsilon)E(t) + \frac{a^2k}{2}\lambda_{\max}(Q) \\ &\quad - 2k(1-\varepsilon)t\lambda_{\min}(Q)\|\dot{w}(t)\| \\ &\quad - \left(\varepsilon - \frac{k}{2a^2}\lambda_{\max}(Q)\right)\|Aw(t)\|^2. \end{aligned}$$

Since a is arbitrary, it can be chosen to be large enough so that $-\frac{k}{2a^2}\lambda_{\max}(Q) > 0$. Therefore, if we put

$$T_2 = \frac{(2-\varepsilon)\|A^{-1/2}\|^2}{2k(1-\varepsilon)\lambda_{\min}(Q)}$$

then

$$\dot{V}(t) \leq 0, \quad \forall t > T_2.$$

Since $E(t)$ is energy dissipating for $t > 0$ and $V(t)$ is also energy dissipating for $t > T_2$, it is easily verified that for $t > T := \max\{T_1, T_2\}$, $E(t)$ can be estimated as

$$E(t) \leq \frac{V(T)}{2(1-\varepsilon)t - C} \leq \frac{[2(1-\varepsilon)T + C]E(0)}{2(1-\varepsilon)t - C},$$

from which we obtain

$$\int_1^\infty E(t)^2 dt \leq \int_1^\infty \frac{[2(1-\varepsilon)T + C]^2 E(0)^2}{[2(1-\varepsilon)t - C]^2} dt < \infty.$$

Noting that $E(t) = \frac{1}{2} \left\| T(t) \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \right\|_{H_2}^2$ and $T(t)$ is a strongly continuous semigroup on H_2 , the above relation implies the exponential stability of solutions of (3), by the well-known equivalent property of L^p ($p \geq 1$)-stability and exponential stability for a strongly continuous semigroup system [14, Theorem 4.1]. Thus we have proved the following theorem.

Theorem 1: The solution of (3) is exponentially stable, provided that Q is a bounded, self-adjoint, and positive definite operator.

Remark 1: It is seen that $V(t)$ is a Lyapunov function for $t > T$.

Remark 2: It should be noted that if a solution is exponentially stable for $t > T$, then it is exponentially stable for $t > 0$. To see this, suppose there exists a $T > 0$, independent of the system initial conditions, such that

$$\begin{aligned} \frac{w(t)}{\dot{w}(t)} &\leq M e^{-\delta(t-T)}, \quad t > T, M \geq 1, \delta > 0. \end{aligned}$$

Then, obviously

$$\frac{w(t)}{\dot{w}(t)} \leq \bar{M} e^{-\delta t}, \quad t > 0$$

where $\bar{M} = M e^{\delta T}$.

Remark 3: It has been shown that under positivity assumption of Q , system (3) generates an analytic semigroup, and moreover all the eigenvalues of (3) lie either on the negative real axis or within a circle with center at $(-\frac{1}{\lambda_{\min}(Q)}, 0)$ and radius $\frac{1}{\lambda_{\min}(Q)}$, on a complex plane [6]. Hence the exponential stability is proved, from a different viewpoint.

III. EXPONENTIAL STABILITY OF (1)

In this section, we investigate the exponential stability of (1) which does not include any natural damping. We solve this problem, in this section, by showing that (1) can be equivalently transformed into a boundary control system well studied in the existing literature [2]. To this end, let us introduce a new variable $y(t, x) = u''(t, x)$. If the initial conditions associated with (1) are sufficiently smooth such that the solution admits continuous spacial derivatives up to sixth order, then taking spacial derivative of both sides of (1) twice yields

$$\ddot{y}(t, x) + y''''(t, x) = 0. \quad (10)$$

Also, from the boundary conditions of (1), it is seen that $\ddot{w}(t, 0) = \ddot{w}'(t, 0) = 0$. Therefore, we obtain the following boundary conditions on $y(t, x)$

$$\begin{aligned} \begin{cases} y(t, 0) = w''(t, 0) = 0 \\ y'(t, 0) = w'''(t, 0) = 0 \\ y''(t, 0) = w''''(t, 0) = -\ddot{w}(t, 0) = 0 \\ y'''(t, 0) = w'''''(t, 0) = -\ddot{w}'(t, 0) - k\dot{w}(t, 0) = -k\dot{y}(t, 0). \end{cases} \end{aligned} \quad (11)$$

At this stage, letting $r = t - x$ and noting that

$$\begin{aligned} \frac{\partial y}{\partial x} &= \frac{\partial y}{\partial r} \frac{\partial r}{\partial x} = -\frac{\partial y}{\partial r}, \quad \frac{\partial^2 y}{\partial x^2} = \frac{\partial^2 y}{\partial r^2} \\ \frac{\partial^3 y}{\partial x^3} &= -\frac{\partial^3 y}{\partial r^3}, \quad \frac{\partial^4 y}{\partial x^4} = \frac{\partial^4 y}{\partial r^4} \end{aligned}$$

we get the following direct velocity feedback boundary control system

$$\begin{aligned} \begin{cases} \ddot{y}(t, r) + y''''(t, r) = 0, \quad r \in (0, t) \\ y(t, 0) = 0, y'(t, 0) = 0 \\ y''(t, 0) = 0 \\ y'''(t, 0) = k\dot{y}(t, 0) \end{cases} \end{aligned} \quad (12)$$

which is well studied by Chen [2] and Morgül [12], and the solution is known to be exponentially stable. Since the solution of (1) can be expressed as $w(t, x) = \int_0^t \int_0^x y(r, s) dr ds$, $w(t, x)$ is also exponentially stable. The transformation from w to y indicates an important relation between direct strain feedback control and direct boundary velocity feedback control, which makes us possible to study sufficient conditions for stability of (2), as will be described below.

IV. A SUFFICIENT CONDITION FOR STABILITY OF (2)

This section is devoted to studying the stability of the hybrid (2). Since it consists of two subsystems highly coupled to each other, one possible method to prove stability of the entire system is to find an appropriate Lyapunov function. Unfortunately, it seems very hard to do this, and until now we have not found an appropriate one. So, we

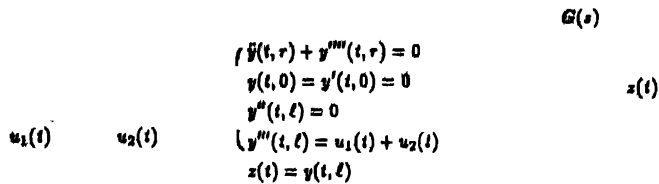


Fig. 1 Block diagram of the transformed closed loop system of (2)

again transform (2) into its corresponding boundary control form and present a sufficient condition for the closed loop stability

Again let $y(t) = u''(t, l)$. By the same way as in the previous section, we can transform (2) into the following form

$$\begin{cases} y(t) + y'''(t) = 0 \\ y(t, 0) = y'(t, 0) = 0 \\ y''(t, l) = 0 \\ y'''(t, l) = -k_1 e(t) - k_2 e(t) + k y(t, l) \\ e(t) + k_1 e(t) + k_2 e(t) = k y(t, l) \end{cases} \quad (13)$$

Now, put

$$z(t) = y(t, l) \quad u_1(t) = -k_1 e(t) - k_2 e(t) \quad u_2(t) = k y(t, l)$$

We obtain a block diagram as shown in Fig. 1. It is interesting to observe that the strain feedback plays a role of an inner loop controller (see the dotted block in Fig. 1) while the motion control part acts as an outer loop controller. Let the transfer function from z to u_1 be denoted by $\bar{K}(s)$. Then

$$\bar{K}(s) = - \frac{k(k_1 s + k_2)}{s^3 + k_1 s + k_2} \quad (14)$$

which is stable and its H^∞ norm $\|\bar{K}(s)\|_\infty$ is easily calculated as

$$\|\bar{K}(s)\|_\infty = \frac{k k_1}{(2\sqrt{k_2^2 + 2k_1^2 k_2^2} + k_1^2 - 2k_2^2 - 2k_1 k_2)^{1/2}} \quad (15)$$

Now, look at the dotted block in Fig. 1 whose transfer function will be denoted by $G(s)$ (note that we do not need the expression of $G(s)$). From the discussion in Section III we know that $G(s)$ is exponentially stable. Hence its H^∞ norm exists.

For direct boundary velocity feedback system

$$\begin{cases} y(t) + y'''(t) = 0 \\ y(t, 0) = y'(t, 0) = 0 \\ y''(t, l) = 0 \\ y'''(t, l) = k y(t, l) + u_1(t) \\ z(t) = y(t, l) \end{cases} \quad (16)$$

there are no existing results on how to calculate the H^∞ norm for the transfer function $G(s)$ from u_1 to z . Here we give an estimation of the upper bound of $\|G\|_\infty$. Taking the inner product of both sides of the first equation in (16) with $y(t, l)$ yields

$$\frac{d}{dt} E(t) + k[y(t, l)]^2 = -u_1(t)y(t, l) \quad (17)$$

where

$$E(t) = \frac{1}{2} \int_0^l ([y(t, l)]^2 + [y''(t, l)]^2) dl$$

represents the total energy stored in (16). Integrating (17) with respect

to t , from 0 to ∞ we obtain

$$k\|z(t)\|_2^2 + \langle u_1(t), z(t) \rangle_2 + E(\infty) - E(0) = 0 \quad (18)$$

where $\langle u_1(t), z(t) \rangle_2 = \int_0^\infty u_1(t) z(t) dt$, $\|z(t)\|_2$ is the induced norm, and

$$F(0) = \frac{1}{2} \int_0^l ([y(0, l)]^2 + [y''(0, l)]^2) dl$$

is the total initial energy. Note that in general $F(\infty)$ is a function of u_1 ; however, $E(\infty) \geq 0$ and for any $\|u_1(t)\|_2 \leq 1$ the upper bound of $F(\infty)$ can be estimated as

$$\begin{aligned} F(\infty) &= E(0) - k\|z(t)\|_2^2 - \langle u_1(t), z(t) \rangle_2 \\ &\leq E(0) - k\|z(t)\|_2^2 + \|z(t)\|_2 \\ &\leq F(0) + \frac{1}{4k} \end{aligned}$$

since the function $f(\|z(t)\|_2) = -k\|z(t)\|_2^2 + \|z(t)\|_2$ attains its maximum value $\frac{1}{4k}$ at $\|z(t)\|_2 = \frac{1}{2k}$. Thus $E(\infty)$ is bounded. So that when u_1 ranges in the unit ball the largest value $\|z(t)\|_2$ satisfying (18) is always smaller than the largest value $\|z(t)\|_2$ satisfying the following inequality

$$k\|z(t)\|_2^2 + \langle u_1(t), z(t) \rangle_2 - F(0) \leq 0 \quad (19)$$

In other words, $\|G(s)\|_\infty$ is small than the largest value $\|z(t)\|_2$ satisfying (19) which will be calculated below.

Let $u_* = \frac{(t)}{\|u_1\|_2}$. It is easy to show that for any u_1 satisfying $\|u_1\|_2 \leq 1$ there holds $\langle u_1, z \rangle_2 \geq \langle u_*, z \rangle_2$ since $\langle u_1, z \rangle_2 \geq -\|u_1\|_2 \|z\|_2 \geq -\|z\|_2 = \langle u_*, z \rangle_2$. So that for any u_1 in the unit ball the largest value that $\|z(t)\|_2$ can achieve is given by the positive root of the following quadratic algebraic equation in $\|z\|_2$

$$k\|z(t)\|_2^2 + \langle u_*(t), z(t) \rangle_2 - F(0) - \|z(t)\|_2 = 0 \quad (20)$$

which is evaluated as

$$\frac{1 + \sqrt{1 + 4kF(0)}}{2k}$$

From this we see that

$$\begin{aligned} \|G(s)\|_\infty &= \sup_{\|u_1\|_2 \leq 1} \|z(t)\|_2 \\ &\leq \frac{1 + \sqrt{1 + 4kF(0)}}{2k} \end{aligned} \quad (21)$$

Since in evaluating transfer function of a linear system the initial conditions are always set to zero we obtain

$$\|G(s)\|_\infty < \frac{1}{k} \quad (22)$$

by setting $E(0) = 0$ in (20). From (15) and (21) we see that

$$\begin{aligned} \|\bar{K}(s)\|_\infty \|G(s)\|_\infty &\leq \frac{k_1^2}{(2\sqrt{k_2^2 + 2k_1^2 k_2^2} + k_1^2 - 2k_2^2 - 2k_1 k_2)^{1/2}} \end{aligned} \quad (23)$$

If the value of the right-hand side of above inequality is less than one then the closed loop stability is implied by the well known small gain theorem. Unfortunately, this is not always the case. To guarantee closed-loop stability however, it is unnecessary to have $\|\bar{K}(s)\|_\infty \|G(s)\|_\infty < 1$ which is quite conservative. Instead, it is sufficient to have

$$|\bar{K}(j\omega)| |G(j\omega)| < 1 \quad (24)$$

for every $\omega \in (-\infty, \infty)$. In general, the low frequency gain of $G(j\omega)$ is very small (near zero) since the output of system (16) is taken to be velocity of $y(t, l)$. On the other hand, the low frequency gain of $\bar{K}(j\omega)$ is approximately one. Thus, if we choose $\sqrt{k_2}$ in such a way that the resonance frequency of $\bar{K}(j\omega)$ is much smaller than that of $G(j\omega)$, then (23) holds for a large range of strain feedback gain k .

It is interesting to note the upper bound $\|K(s)\|_\infty \|G(s)\|_\infty$ of $|K(j\omega)| |G(j\omega)|$ is independent of k , which is particularly useful for feedback design.

Remark 4: For the closed-loop system shown in Fig. 1, if $\|G\|_\infty \|K\|_\infty < 1$, then the closed-loop system is internally exponentially stable [8, Lemma 2.3], since G and K are internally exponentially stable [15]. Under the present condition $|G(j\omega)| |K(j\omega)| < 1$, whether input-output stability implies internal stability remains a problem.

V. CONCLUSION

Some newly obtained stability results have been presented in this paper. Specifically,

- 1) For a general abstract differential equation on a Hilbert space, exponential stability was shown under an additional condition (this condition is equivalent to the existence of a Voigt damping term in dynamic model in the case of control of flexible arms);
- 2) For a DSFB controlled closed-loop equation in partial differential equation form, the exponential stability is verified, which clearly indicates the power of direct strain feedback control of flexible arms and reveals an important relationship between the DSFB control and the energy dissipating direct boundary velocity feedback control;
- 3) For a simultaneous vibration/motion controlled hybrid equation, a sufficient condition for the entire system to be stable is given, which can be used as a guideline in choosing feedback gains in (2).

ACKNOWLEDGMENT

The first author thanks Prof. Y. Inouye for some useful discussions. The authors would also like to thank the anonymous referees for helpful comments.

REFERENCES

- [1] R. H. Cannon, Jr. and E. Schmitz, "Initial experiments on the end-point control of a flexible one-link robot," *Int. J. Robotics Res.*, vol. 3, no. 3, pp. 62-75, 1984.
- [2] G. Chen, M. C. Delfour, A. M. Krall, and G. Payre, "Modeling, stabilization and control of serially connected beams," *SIAM J. Contr. Optim.*, vol. 25, pp. 526-546, 1987.
- [3] G. Chen and D. L. Russel, "A mathematical model for linear elastic systems with structural damping," *Quarterly Appl. Math.*, pp. 433-454, 1982.
- [4] R. F. Curtain and A. J. Pritchard, *Functional Analysis in Modern Applied Mathematics*. New York: Academic, 1977.
- [5] J. H. Davis and R. M. Hirschorn, "Tracking control of flexible robot link," *IEEE Trans. Automat. Contr.*, vol. 33, no. 3, pp. 238-248, 1988.
- [6] B. Z. Guo and Y. Luo, "Semigroup approach to the stability of a direct strain feedback control system of elastic vibration with structure damping," *Appl. Math. Letters*, vol. 7, no. 1, pp. 95-100, 1994.
- [7] F. Huang, "On the mathematical model for linear elastic systems with analytic damping," *SIAM J. Contr. Optim.*, vol. 26, no. 3, pp. 714-724, 1988.
- [8] B. Van Keulen, "A state-space approach to H_∞ -control problems for infinite-dimensional systems," in *Analysis and Optimization of Systems: State and Frequency Domain Approaches for Infinite Dimensional Systems*. R. F. Curtain, et al., Ed. New York: Springer-Verlag, 1993.
- [9] Z. H. Luo, "Stability analysis of direct strain feedback control of flexible robot arms," in *Proc. Amer. Contr. Conf.*, 1992, pp. 3319-3323.
- [10] —, "Direct strain feedback control of flexible robot arms: new theoretical and experimental results," *IEEE Trans. Automat. Contr.*, vol. 38, no. 11, pp. 1610-1622, 1993.
- [11] —, "On A-dependent operators and the associated abstract equations," in *Proc. 31st IEEE Conf. Decis. Contr.*, 1992, pp. 3119-3120.

- [12] Ö. Morgül, "Dynamic boundary control of a Euler-Bernoulli beam," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 639-642, 1992.
- [13] —, "Orientation and stabilization of a flexible beam attached to a rigid body: Planar motion," *IEEE Trans. Automat. Contr.*, vol. 36, no. 8, pp. 953-962, 1991.
- [14] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*. New York: Springer-Verlag, 1983.
- [15] R. Rebarber, "Conditions for the equivalence of internal and external stability for distributed parameter systems," *IEEE Trans. Automat. Contr.*, vol. 38, no. 6, pp. 994-998, 1993.
- [16] Y. Sakawa, F. Matsuno, and S. Fukushima, "Modeling and feedback control of a flexible one-link robot," *J. Robotic Syst.*, vol. 2, pp. 453-472, 1985.

H^∞ Optimal and Suboptimal Controllers for Infinite Dimensional SISO Plants

Onur Toker and Hitay Özbay

Abstract—This paper presents a simple formula for \mathcal{H}^∞ optimal and suboptimal controllers for unstable SISO distributed plants, with rational weighting functions. The controller is expressed in terms of i) inner and outer parts of the plant, ii) a finite dimensional spectral factor obtained from the weighting functions, and iii) a rational function satisfying certain interpolation conditions. Under certain genericity assumptions, this rational function is of dimension less than or equal to $n_1 + l - 1$ ($n_1 + l$ in the suboptimal case), where l is the number of unstable poles of the plant and n_1 is the order of the sensitivity weighting function. There are $2(n_1 + l)$ ($2(n_1 + l + 1)$ in the suboptimal case) linear equations, which determine this rational function. These linear equations can be written directly from the structure of the controller.

NOTATION

\mathbb{R}	The set of real numbers.
\mathbb{C}	The set of complex numbers.
\mathbb{C}_+	$\{s \in \mathbb{C} : \operatorname{Re}(s) > 0\}$.
\mathbb{D}	Open unit disc, $\{z \in \mathbb{C} : z < 1\}$.
\mathbb{T}	Unit circle, $\{z \in \mathbb{C} : z = 1\}$.
\mathcal{L}^∞ , $\mathcal{L}^\infty(\mathbb{T})$	Banach space of essentially bounded functions on $j\mathbb{R}$, \mathbb{T} .
\mathcal{H}^∞ , $\mathcal{H}^\infty(\mathbb{D})$	\mathcal{L}^∞ functions which admit bounded analytical extensions to \mathbb{C}_+ , \mathbb{D} .
\mathcal{RH}^∞	Real rational functions in \mathcal{H}^∞ .
\mathcal{L}^2 , $\mathcal{L}^2(\mathbb{T})$	Hilbert space of square integrable functions on $j\mathbb{R}$, \mathbb{T} .
\mathcal{H}^2 , $\mathcal{H}^2(\mathbb{D})$	\mathcal{L}^2 functions which admit analytical extensions to \mathbb{C}_+ , \mathbb{D} .
$\mathcal{P}_{\mathcal{M}}$	The orthogonal projection onto a subspace \mathcal{M} of \mathcal{L}^2 .
	The orthogonal complement of \mathcal{H}^2 in \mathcal{L}^2 .
\mathcal{B}	The unit ball of \mathcal{H}^∞ .
\mathbf{F}	The reflection (or flip) operator on $\mathcal{L}^2(\mathbb{T})$, $\mathbf{F}f(z) = z^{-1}f(z^{-1})$.
	The shift operator on $\mathcal{H}^2(\mathbb{D})$, $\mathbf{S}f(z) = zf(z)$.

Manuscript received February 22, 1994; revised June 14, 1994. This work was supported in part by NSF Grant MSS-9203418 and AFOSR Grant F49620-93-1-0288.

The authors are with the Department of Electrical Engineering, The Ohio State University, Columbus, OH 43210 USA.
IEEE Log Number 9408796.

Γ_v The Hankel operator with symbol v , $\Gamma_v = \mathbf{P}_{\mathcal{H}_1^2} v$.
 $F|_I = 0$ $F(s) = 0$ whenever $f(s) = 0$.

I. INTRODUCTION

In this paper the mixed sensitivity minimization (two-block) problem is considered for a class of infinite dimensional SISO, LTI plants with finitely many unstable poles, with rational weighting functions. The purpose of this paper is to present a formula for \mathcal{H}^∞ optimal and suboptimal controllers. In the one block problem (sensitivity minimization) the \mathcal{H}^∞ optimal performance and controller can be computed by finding the largest singular value, and corresponding singular vector, of an infinite rank Hankel (or "skew Toeplitz") operator. Earlier studies on the one-block problem, for stable infinite dimensional plants, have shown that the computation of singular values and vectors of this infinite rank operator reduces to finding a solution to a set of finitely many linear equations, [12], [5]. In the two block problem, the operator we are dealing with is a "Hankel plus Toeplitz" type, with possibly infinite rank Hankel part. This problem can be reduced to a set of finitely many linear equations; see [7] for the finite dimensional case and [13] for infinite dimensional stable plants case. More recently, these results have been extended to the case where the infinite dimensional plant has finitely many unstable modes; see [8] and [2] for detailed reviews and further references. This paper extends these results to the suboptimal case and gives a simple formula for the optimal controller as well as all suboptimal controllers. Previously it was known (see, e.g., [10], [8] and references therein) how to compute the optimal controller "numerically," by using Youla parameterization and computing the singular values and vectors of the associated infinite rank operator. We follow a similar technique to solve a more general problem. Moreover, we obtain a simple closed form expression for the \mathcal{H}^∞ controllers in terms of inner outer factors of the plant, a spectral factor of the weights, and a rational function to be determined. The set of linear equations, which determines the optimal performance and this rational function, can be written directly from the controller structure.

The rest of this paper is organized as follows. Two block \mathcal{H}^∞ optimal and suboptimal control problems are defined in Section II. Main results are presented in Section III. The proofs can be found in Section IV, and concluding remarks are made in Section V.

II. \mathcal{H}^∞ OPTIMAL AND SUBOPTIMAL MIXED SENSITIVITY

In this paper the plant and the controller are represented by their transfer functions $P(s)$ and $C(s)$, respectively. A system, whose transfer function is $G(s)$, will be said to be stable if $G \in \mathcal{H}^\infty$. Throughout the paper we consider "real" functions only, i.e., $\bar{R}(s) = R(\bar{s})$.

This paper deals with the mixed sensitivity reduction/minimization problem for a given plant $P(s)$ and two rational weighting functions $W_1(s)$ and $W_2(s)$.

Assumptions on the Problem Data: We assume that the plant admits a coprime factorization of the form $P(s) = m_n(s)N_0(s)/m_d(s)$ where $m_d(s) = \prod_{k=1}^l \frac{s - \alpha_k}{s + \alpha_k}$, $\alpha_1, \dots, \alpha_l \in \mathbb{C}_+$, are distinct, and $m_n \in \mathcal{H}^\infty$ is inner (i.e., stable all-pass function, possibly infinite dimensional) and $N_0 \in \mathcal{H}^\infty$ is outer (i.e., stable minimum phase, possibly infinite dimensional). Moreover, the weights are assumed to be rational with $W_1(s)$ being nonconstant and $W_1, (W_2 N_0), (W_2 N_0)^{-1} \in \mathcal{H}^\infty$; see [8] for a detailed discussion on these assumptions, and relations to other types of two block \mathcal{H}^∞ control problems.

The optimal \mathcal{H}^∞ performance is defined by

$$\gamma_0 := \inf_{C \text{ stabilizes } P} \left\| \begin{bmatrix} W_1 S \\ W_2 T \end{bmatrix} \right\|_\infty$$

where $S = (1 + PC)^{-1}$ and $T = PC(1 + PC)^{-1}$ are the sensitivity and complementary sensitivity functions. The condition " C stabilizes P " means that closed-loop transfer functions S , CS , and PS belong to \mathcal{H}^∞ . The optimal \mathcal{H}^∞ controller, denoted by C_{opt} , is the one which stabilizes the plant P , and yields

$$\left\| \begin{bmatrix} W_1(1 + PC_{\text{opt}})^{-1} \\ W_2 PC_{\text{opt}}(1 + PC_{\text{opt}})^{-1} \end{bmatrix} \right\|_\infty = \gamma_0.$$

The suboptimal \mathcal{H}^∞ control problem is to parameterize the set

$$\mathcal{C}_\rho = \left\{ C: C \text{ stabilizes } P, \left\| \begin{bmatrix} W_1 S \\ W_2 T \end{bmatrix} \right\|_\infty \leq \rho \right\} \quad (1)$$

for a given suboptimal performance level $\rho > \gamma_0$.

III. OPTIMAL AND SUBOPTIMAL \mathcal{H}^∞ CONTROLLERS

The main results of the paper are stated in this section. Their proofs are given in Section IV.

Let $\eta_1, \dots, \eta_{n_1} \in \mathbb{C}_+$, $n_1 \geq 1$, be the poles of $W_1(-s)$; if η_i has multiplicity ℓ_i then it is assumed to be repeated ℓ_i times in this list. The zeros of

$$F_\rho(s) := \left(\frac{W_1(-s)W_1(s)}{\rho^2} - 1 \right) \quad (2)$$

are denoted by $\beta_1, \dots, \beta_{2n_1}$, and they are assumed to be distinct for the given ρ . Then, β_i 's can be enumerated in such a way that $\beta_1, \dots, \beta_{n_1}$ are in \mathbb{C}_+ , and $\beta_{n_1+i} = -\beta_i$. Now define

$$G_\rho(s) := G_{\text{opt}}(s) \prod_{k=1}^{n_1} \frac{s - \eta_k}{s + \eta_k} \quad (3)$$

where $G_\rho \in \mathcal{H}^\infty$ is minimum phase and determined from the spectral factorization

$$G_\rho(s)G_\rho(-s) = \frac{W_1(-s)W_1(s)}{\rho^2} - 1 = \frac{W_2(-s)W_2(s)}{\rho^2} - 1 \quad (4)$$

Genericity Assumptions For the given $\rho > \gamma_0$ we assume that β_i 's are distinct and none of them coincide with any of the α_i 's. Furthermore, β_i 's are such that $m_n(\beta_i) \neq 0$, for all $i = 1, \dots, 2n_1$. We also assume that $\gamma_0 > \gamma_{\min}$, where γ_{\min} is the smallest ρ such that the right-hand side of (4) is greater or equal to zero for all $s = j\omega$.

Now, to obtain a parameterization of \mathcal{C}_ρ , pick an arbitrary real number $a > 0$. Then, we have the following formula.

Theorem 1: Consider the suboptimal \mathcal{H}^∞ control problem defined in Section II, and suppose that the above genericity assumptions hold. Then, $C_{\text{subopt}} \in \mathcal{C}_\rho$ if and only if

$$C_{\text{subopt}}(s) = E_\rho(s)m_d(s) \frac{N_0(s)^{-1}F_\rho(s)L_l(s)}{1 + m_n(s)F_\rho(s)L_l(s)} \quad (5)$$

where

$$L_l(s) = \frac{L_2(s) + L_1(-s)U(s)}{L_1(s) + L_2(-s)U(s)}, \quad \text{for some } U \in \mathcal{B}$$

and $L_1(s), L_2(s)$ are polynomials of degree $\leq n_1 + l$ satisfying interpolation conditions

$$0 = L_1(\beta_k) + m_n(\beta_k)F_\rho(\beta_k)L_2(\beta_k) \quad k = 1, \dots, n_1 \quad (6)$$

$$0 = L_1(\alpha_k) + m_n(\alpha_k)F_\rho(\alpha_k)L_2(\alpha_k) \quad k = 1, \dots, l \quad (7)$$

$$0 = L_2(-\beta_k) + m_n(\beta_k)F_\rho(\beta_k)L_1(-\beta_k) \quad k = 1, \dots, n_1 \quad (8)$$

$$0 = L_2(-\alpha_k) + m_n(\alpha_k)F_\rho(\alpha_k)L_1(-\alpha_k) \quad k = 1, \dots, l \quad (9)$$

$$0 = L_2(-a) + (E_\rho(a) + 1)F_\rho(a)m_n(a)L_1(-a) \quad (10)$$

$$1 = L_1(-a). \quad (11)$$

In the above system of equations, $\alpha > 0$ is arbitrary. For different values of α one can obtain different parameterizations of the suboptimal controllers.

Remark 1 The genericity assumption $\gamma_0 > \gamma_{\min}$ is actually not needed for Theorem 1, but it will be necessary for Theorem 2 which deals with the optimum case. Furthermore, in the optimum case the genericity assumptions are assumed to hold for $\rho = \gamma_0$.

Remark 2 If $W_1(s) = n(s)/d(s)$ with $n(s), d(s) \in \mathbb{R}[s]$ ($n(s), d(s) = 1$), $W_1^{-1} \in \mathcal{H}^\infty$ then $\Lambda_1(s) = n(s)n(-s)$ and $D_1(s) = d(s)d(-s)$ are relatively prime. If $\Lambda(s) - \rho^2 D_1(s)$ has a multiple root then at that point $\Lambda' - \rho^2 D_1' = 0$ and hence $D_1 \Lambda' - \Lambda D_1' = 0$. But $D_1 \Lambda' - \Lambda D_1'$ is not identically equal to zero, and hence has only finitely many zeros, so there are at most finitely many ρ values for which the genericity assumption on β is not satisfied. The fact that $D_1 \Lambda' - \Lambda D_1'$ is not identically equal to zero can be shown as follows. Let $\Lambda = Y$ be polynomials such that $\Lambda \Lambda' + Y_1 D_1 = 1$ and assume that $D_1 \Lambda' - \Lambda D_1' = 0$. Then we get $\Lambda_1 D_1 \Lambda' - \Lambda_1 \Lambda D_1' = 0$, $D_1' \Lambda \Lambda' + D_1 Y_1 D_1' = D_1'$ and hence D_1 divides D_1' which is absurd. This contradiction proves that $D_1 \Lambda' - \Lambda D_1'$ is not identically equal to zero, and hence shows that there are at most finitely many ρ value for which the genericity assumption on β is not satisfied.

Theorem 2 Suppose that C_{\min} is unique and γ_0 is the largest singular value of the associated Hankel plus Toeplitz operator defined in the next section. Moreover, suppose that the above genericity assumptions hold for γ_0 . Then

$$C(s) = F(s)m_1(s) \frac{\Lambda_1(s) - I_1(s)I_1(s)}{1 + m_1(s)F_0(s)I_1(s)} \quad (12)$$

where $I = I_1/L_1$ and $I_1(s), I_1(s)$ are nonzero polynomials with degrees less than or equal to $(n_1 + l - 1)$ satisfying (6)–(9) with $\rho = \gamma_0$.

To find C we first find an upper bound, e.g., by taking any stabilizing controller and evaluating its performance. Then we replace ρ with a new parameter in (6)–(9). Note that (6)–(9) can be written as $M\Psi = 0$ where Ψ is a $2(n_1 + l)$ column vector formed by the coefficients of $I_1(s)$ and $I_1(s)$ and M is a square matrix which depends on the parameter ρ . Therefore γ_0 can be found by plotting the smallest singular values of M as ρ varies between an upper and lower bound. The largest value of ρ for which the plot shows a zero is γ_0 (see the below example). In this sense, the algorithm of finding γ_0 is similar to standard iteration ([3], [6], [7]). But here the size of the matrix whose singularity is to be determined is $2(n_1 + l) \times 2(n_1 + l)$ even though the underlying system is infinite dimensional.

Example Let $P(s) = e^{-1}/(s-1)$, $W_1(s) = 2(s+1)/(10s+1)$ and $W_2(s) = 0.2/(s+1)$. In this example $m_1(s) = e^{-1}/(s-1)$, $m_2(s) = (s-\alpha_1)/(s+\alpha_1)$ and $\Lambda_1(s) = 1/(s+L_1)$ where $\alpha_1 = 1$. Note that we need to find $\gamma_0 = F$ and $I(s) = L_1(s)/I_1(s)$ in the controller expression given by (12). Since $n_1 = 1$ and $l = 1$ polynomials $I_1(s)$ and $I_1(s)$ are both first order. Therefore there are four unknown coefficients $\Psi = [I_{10} \ L_{20} \ L_{11} \ L_{11}]^T$ where $L_1(s) = I_{11}s + I_{10}$ and $L_2(s) = L_{21}s + L_{20}$. When $h = 0.2$ it can be shown that $0.2 < \gamma_0 < 1.5$ (see [4] and [9]).

In this case the largest value of γ which makes M_γ singular is $\gamma_0 = 0.6819$. It is obtained from Fig. 1 where the smallest singular value of M_γ is plotted. Once γ_0 is obtained from this plot, a nonzero Ψ_0 , which satisfies $M_{\gamma_0}\Psi_0 = 0$ can be obtained easily. The entries of Ψ_0 give $I(s) = L_2(s)/I_1(s)$ which determine the optimal controller via (12). For the above example $I(s)$ and F_{γ_0} can be computed as

$$I(s) = \frac{s+0.2129}{s-0.2129}, \quad F_{\gamma_0}(s) = \frac{3.566(s-0.1)}{(s+1.265)(s+0.788)}$$

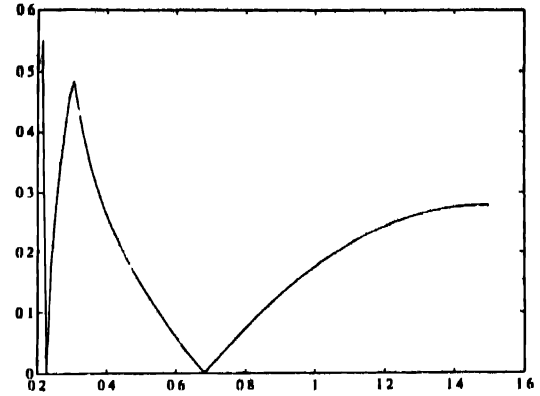


Fig. 1 $\sigma_{\min}(M_\gamma)$ versus γ

For the suboptimal performance level $\rho = 0.7$ we get $F_\rho(s) = \frac{1}{(s+1)} \frac{2(s-1)}{(s+0.2129)}$ and if we choose $\alpha = 2$ then

$$F_\rho(s) = L_{\rho, \alpha}(s) \frac{1 + I(s)/D(s)}{1 + D(s)I(s)}$$

parameterizes all suboptimal controllers via (5) where

$$L_{\rho, \alpha}(s) = 0.96 \frac{(s+1.999)(s+0.2146)}{(s+1.94)(s-0.208)} \\ D(s) = 0.96 \frac{(s-1.999)(s-0.2146)}{(s+1.94)(s-0.208)}$$

We see that for this example $\sigma_{\min}(M_\gamma)$ behaves numerically well, i.e., it is easy to detect γ_0 from Fig. 1. Also note that in the optimal case $I(s)$ is all pass (but possibly unstable) and when ρ is close to γ_0 , $D(s)$ is close to being inner and $L_{\rho, \alpha}(s)$ and $I_\rho(s)$ are close to $I(s)$ of the optimal case.

IV. PROOFS OF THE MAIN RESULTS

Consider the \mathcal{H}^∞ control problems defined in Section II and define $G_r = \Lambda/m \in \mathcal{H}^\infty$ and $G_l = m/l \in \mathbb{RH}^\infty$ then $P = G_r/G_l$ is a coprime factorization. Let $\Lambda, Y \in \mathcal{H}^\infty$ be such that $\Lambda G_r + Y G_l = 1$. All stabilizing controllers are in the form (see [11])

$$C = \frac{\Lambda + Q G_l}{Y - Q G_r} \quad (13)$$

where Q is a free parameter in \mathcal{H}^∞ . Following the notation of [9], define G to be such that $W_1 W_1^* + W_2 W_2^* = G_r G_r^*$ and $G^{-1} \in \mathcal{H}^\infty$. Let $b_1 = b_1 G_r^* W_1 W_1^* \in \mathbb{RH}^\infty$. Then let $W_0 \in \mathbb{RH}^\infty$ be such that $\Lambda = (G_r - \Lambda W_0)/G_l$ and set $Q_1 = G_l W_0 Q + \Lambda$. $G_r = W_1 W_1^* G^{-1} = b_1 m_1 m_1^* = b_1 m_m$. It was shown that [9] the optimal control problem reduces to the computation of

$$\gamma_0 = \inf_{Q_1 \in \mathcal{H}^\infty} \left\| \frac{W_0 - m_1 W_0 - m Q_1}{G_r} \right\|_\infty \quad (14)$$

and determination of Q_{1, γ_0} and the suboptimal control problem reduces to

$$\{Q_1 \in \mathcal{H}^\infty \mid \left\| \frac{W_0 - m_1 W_0 - m Q_1}{G_r} \right\|_\infty \leq \gamma\} \quad (15)$$

The optimal performance γ_0 is the norm of the so-called "Hankel plus Toeplitz" operator $[\Gamma_{\rho(m_1 W_0 - W_0)} \Psi]_{\rho, 2}$ where Ψ_ρ denotes the Toeplitz operator with symbol G_r [8]. Note that $W_0, W_0 - m_1 W_0, G_r$ are stable and rational, and m_1 is inner and possibly infinite dimensional. Now let f be an \mathcal{H}^∞ function such that $f f^* = \gamma^2 - G_r G_r^*$ and $f^{-1} \in \mathcal{H}^\infty$. Note that by genericity assumptions $\gamma_0 > \gamma_{\min}$, and it is easy to show that $\gamma_{\min} = \|G_r\|_\infty$. Define $u_\gamma = W_0/f_\gamma$

$\hat{u}_\gamma = \hat{W}_0/f_\gamma$, and $Q_2 = f_\gamma^{-1}Q_1$. Then γ_0 is the maximum $\gamma > \gamma_{\min}$ value such that one is a singular value of the Hankel operator $\Gamma = \Gamma_{m^*(u_\gamma - m_1 \hat{u}_\gamma)}$, (see, e.g., [8] and references therein), furthermore

$$m^*(u_{\gamma_0} - m_1 \hat{u}_{\gamma_0}) - Q_{2, \text{opt}} = \frac{y_0^*}{x_0} \quad (16)$$

where x_0, y_0 are nonzero \mathcal{H}^2 functions satisfying

$$\Gamma x_0 = y_0^*, \quad \Gamma^* y_0^* = x_0^*. \quad (17)$$

Claim: When $\gamma > \gamma_0$ we have

$$m^*(u_\gamma - m_1 \hat{u}_\gamma) - Q_{2, \text{subopt}} = \frac{y_0^* + z_0 x_0^* U^*}{x_0 + z_0 y_0 U^*}, \quad \begin{matrix} s-a \\ s+a \end{matrix}, \quad U \in \mathcal{B} \quad (18)$$

where x_0, y_0 are \mathcal{H}^2 functions satisfying

$$y_0^* = \Gamma x_0, \quad x_0 = \frac{1}{s+a} + \Gamma^* y_0^*. \quad (19)$$

Proof. Let $V_\gamma(s) = m^*(s)(u_\gamma(s) - m_1(s)\hat{u}_\gamma(s))$ and $v_\gamma(z) = V_\gamma(a \frac{1+z}{1-z})$, for an arbitrary $a > 0$. Note that $\mathcal{S}_\gamma = \{f_\gamma Q_2 \in \mathcal{H}^\infty : \|V_\gamma - Q_2\|_\infty \leq 1\}$. So, to find all such Q_2 's one can find all $q_2 \in \mathcal{H}^\infty(\mathbb{D})$ such that $\|v_\gamma - q_2\|_\infty \leq 1$ and set $Q_2(s) = q_2(\frac{s-a}{s+a})$. The set of all such q_2 's is given by [1]

$$v_\gamma - q_2 = \frac{\mathcal{E}p^* + q^*}{p + \mathcal{E}q}, \quad \mathcal{E} \in \mathcal{B}(\mathbb{D})$$

where $p, q \in \mathcal{H}^2(\mathbb{D})$ such that $(I - \mathbf{F}\Gamma_v \Gamma_v^* \mathbf{F})p = 1$, and $q = \mathbf{S}\Gamma_v^* \mathbf{F}p$ (we use $v = v_\gamma$ whenever the notation is clear from the context).

Set $x(z) = p(z)$, $y(z) = z^{-1}q(z)$, then we have

$$y = \Gamma_v^* \mathbf{F}x, \quad \text{and} \quad x = 1 + \mathbf{F}\Gamma_v y.$$

By defining $x_0(s) = \frac{1}{s+a}x(\frac{s-a}{s+a})$, $y_0(s) = -\frac{1}{s+a}y(\frac{s-a}{s+a})$, we obtain an equivalent set of equations in terms of $x_0, y_0 \in \mathcal{H}^2$

$$y_0^* = \Gamma x_0, \quad x_0 = \frac{1}{s+a} + \Gamma^* y_0^*. \quad (20)$$

Finally, by setting $U(s) = -\mathcal{E}(\frac{s-a}{s+a})$, we obtain (18). \square

In summary, we consider the following system of equations, to find C_{opt} and C_p

$$x_0, y_0 \in \mathcal{H}^2, \quad y_0^* = \Gamma x_0, \quad x_0 = \frac{1}{s+a} + \Gamma^* y_0^*. \quad (21)$$

The suboptimal control problem is equivalent to (21) for $\delta = 1$, i.e., one has to find the unique nonzero x_0, y_0 satisfying (21), and then (18) can be used to obtain all $Q_{2, \text{subopt}}$. The optimal control problem amounts to solving (21) for $\delta = 0$, i.e., one has to find nonzero x_0, y_0 satisfying (21) and then $Q_{2, \text{opt}}$ can be obtained from (18) by setting $U = 0$.

A. On the Solution of x_0 and y_0

Let $u_\gamma = b_\gamma/k_\gamma$, $m_d = A^*/A$ where $b_\gamma, k_\gamma, A \in \mathbb{R}[s]$, $(b_\gamma, k_\gamma) = 1$. Define $n = \deg(k_\gamma)$ and $l = \deg(A)$. It is clear that (21) is equivalent to (22)–(24)

$$x_0, y_0, \phi_1, \phi_2 \in \mathcal{H}^2, \quad (22)$$

$$y_0^* = m^*(u_\gamma - m_1 \hat{u}_\gamma)x_0 + \phi_1, \quad (23)$$

$$x_0 = \frac{\delta}{s+a} + m(u_\gamma^* - m_1^* \hat{u}_\gamma^*)y_0^* + \phi_2^*, \quad (24)$$

Note that (23) implies that $m y_0^* \in \mathcal{H}_+^2$ and (24) implies that $m^*(x_0 - \frac{\delta}{s+a}) \in \mathcal{H}_+^2$, i.e., $m^*x_0 - \frac{\delta m(a)}{s+a} \in \mathcal{H}_+^2$. Since

$$\begin{aligned} \phi_1 + (u_\gamma - m_1 \hat{u}_\gamma) \frac{\delta m(a)}{s+a} \\ = -\mathbf{P}_{\mathcal{H}^2}(u_\gamma - m_1 \hat{u}_\gamma) \left(m^*x_0 - \frac{\delta m(a)}{s+a} \right) \\ = -\mathbf{P}_{\mathcal{H}^2} u_\gamma \left(m^*x_0 - \frac{\delta m(a)}{s+a} \right) \\ + \mathbf{P}_{\mathcal{H}^2} m_d^* \hat{u}_\gamma x_0 - \mathbf{P}_{\mathcal{H}^2} m_1 u_\gamma \frac{\delta m(a)}{s+a} \end{aligned}$$

we have

$$\begin{aligned} \mathbf{P}_{\mathcal{H}^2} u_\gamma \left(m^*x_0 - \frac{\delta m(a)}{s+a} \right) &= \frac{p_{1,1}}{k_\gamma} \\ \mathbf{P}_{\mathcal{H}^2} m_d^* \hat{u}_\gamma x_0 &= m_d^* \hat{u}_\gamma x_0 - \frac{p_{1,2}}{A^*} \end{aligned}$$

where $p_{1,1}, p_{1,2} \in \mathbb{R}[s]$ with $\deg(p_{1,1}) \leq n-1$, $\deg(p_{1,2}) \leq l-1$. Hence we obtain

$$\phi_1 = \frac{p_{1,1}}{k_\gamma A^*} + m_d^* \hat{u}_\gamma x_0 - u_\gamma \frac{\delta m(a)}{s+a}$$

where $p_1 \in \mathbb{R}[s]$ and $\deg(p_1) \leq n+l-1$. Similarly, since

$$\begin{aligned} \phi_2 &= -\mathbf{P}_{\mathcal{H}_+^2} m(u_\gamma^* - m_1^* \hat{u}_\gamma^*)y_0^* \\ &= -\mathbf{P}_{\mathcal{H}_+^2} m u_\gamma^* y_0^* - \mathbf{P}_{\mathcal{H}_+^2} m_d \hat{u}_\gamma^* y_0^* \end{aligned}$$

we have

$$\begin{aligned} \mathbf{P}_{\mathcal{H}_+^2} m u_\gamma^* y_0^* &= \frac{p_{2,1}}{k^*} \\ \mathbf{P}_{\mathcal{H}_+^2} m_d \hat{u}_\gamma^* y_0^* &= m_d u_\gamma^* y_0^* - \frac{p_{2,2}}{A} \end{aligned}$$

for some $p_{2,1}, p_{2,2} \in \mathbb{R}[s]$ with $\deg(p_{2,1}) \leq n-1$, $\deg(p_{2,2}) \leq l-1$. Thus

$$\phi_2 = \frac{p_{2,2}}{k^* A} + m_d u_\gamma^* y_0^*$$

for some $p_2 \in \mathbb{R}[s]$ and $\deg(p_2) \leq n+l-1$. Now define $P_1(s) = (s+a)p_1(s) - \delta b_\gamma(s)A^*(s)m(a)$ and $P_2(s) = (s+a)p_2(s) + \delta k_\gamma^*(s)A(s)$. Then, we have

$$x_0 = -\frac{\kappa_\gamma x_2 + m_1 v_\gamma x_1}{(s+a)\pi A}, \quad (25)$$

$$y_0^* = -\frac{k_\gamma^* P_1 + m_1^* b_\gamma P_2}{(s+a)\pi A^*}, \quad (26)$$

$$\phi_1 = \frac{P_1 \pi - k_\gamma u_\gamma (k_\gamma^* P_1 + m_1^* b_\gamma P_2)}{(s+a)\pi A^* k_\gamma}, \quad (27)$$

$$\phi_2^* = \frac{P_2 \pi - k_\gamma^* \hat{u}_\gamma^* (k_\gamma^* P_1 + m_1^* b_\gamma P_2)}{(s+a)\pi A L^*}, \quad (28)$$

where P_1 and P_2 satisfy

$$\deg(P_1) \leq n+l, \quad \deg(P_2) \leq n+l \quad (29)$$

$$k_\gamma P_2 + m_1 b_\gamma^* P_1|_\pi = 0 \quad (30)$$

$$P_2 \pi - k_\gamma^* \hat{u}_\gamma^* (k_\gamma^* P_1 + m_1^* b_\gamma P_2)|_{A^*} = 0 \quad (31)$$

$$P_2 \pi - k_\gamma^* u_\gamma^* (k_\gamma^* P_1 + m_1^* b_\gamma P_2)|_A = 0 \quad (32)$$

$$k_\gamma^* P_1 + m_1^* b_\gamma P_2|_{s=-a} = 0 \quad (33)$$

$$P_2(-a) = -\delta b_\gamma A^* m|_{s=-a}. \quad (34)$$

Conversely, if (29)–(34) hold, we can define x_0, y_0, ϕ_1, ϕ_2 by (25)–(28). In this case (22)–(24) hold and hence (21) holds.

If $\delta = 0$, (33) and (34) imply that $P_1(-a) = P_2(-a) = 0$. In this case (30)–(32) hold for P_i replaced by P'_i , $i = 1, 2$ and (29) replaced by $\deg(P'_i) \leq n+l-1$. Furthermore, (33)–(34) are redundant and

(25)–(28) hold if P_i replaced by P_i^* , $i = 1, 2$ and the $(s + a)$ factor is cleared from the denominators. Since (16) depends only on y_0^*/x_0 , it is enough to solve (21) for $\delta = 0$ and to set $U = 0$ in (18) to solve the optimal problem.

B. On the Structure of the Controller

Let $W_1 = B_1/K_1$, $W_2 = B_2/K_2$ with $(B_1, K_1) = (B_2, K_2) = 1$. Define A_{C_i} to be a stable polynomial such that

$$A_{C_i} A_{C_i}^* = B_1 B_1^* K_2 K_2^* + B_2 B_2^* K_1 K_1^*. \quad (35)$$

Then, it is easy to see that $G = \frac{A_G}{K_1 K_2}$. Similarly define B_γ to be a stable polynomial such that

$$B_\gamma B_\gamma^* = \gamma^2 A_{C_i} A_{C_i}^* - B_1 B_2 B_1^* B_2^*. \quad (36)$$

Then, we can define $f_\gamma = \frac{B_\gamma}{A_{C_i}}$. So G_γ defined in Section II is equal to $G_\gamma = \frac{\gamma^2 K_1 K_2}{B_\gamma}$. On the other hand, since b_1 is the Blaschke product of minimal order such that $W_0 = b_1 G^{-*} W_1 W_1^* = b_1 \frac{K_2^* B_1 B_1^*}{A_{C_i}^* K_1} \in \mathcal{H}^\infty$, we obtain $b_1 = \frac{A_{C_i}^*}{A_G}$. Furthermore, note that $W_0 = \frac{B_1 B_1^* K_2^*}{K_1 A_G}$, $u_\gamma = \frac{B_1 B_1^* K_2^*}{K_1 B_\gamma}$, $b_\gamma = B_1 B_1^* K_2^*$, $k_\gamma = K_1 B_\gamma$. Since $\pi(s) = b_\gamma b_\gamma^* - k_\gamma k_\gamma^* = (B_1 B_1^* - \gamma^2 K_1 K_1^*) A_{C_i} A_{C_i}^*$, and $m_1 = m_n \frac{A_G}{A_{C_i}}$, (30) implies that

$$P_1 = A_{C_i} L_2, \quad P_2 = A_{C_i}^* L_1. \quad (37)$$

In this case (30) reduces to

$$L_1 + m_n F_\gamma L_2|_{E_\gamma} = 0 \quad (38)$$

because $u_\gamma = \frac{W_1 W_1^*}{\gamma^2} F_\gamma$ and hence $(u_\gamma - F_\gamma)|_{E_\gamma} = 0$.

Similarly, (31)–(32) reduce to

$$L_1 + m_n F_\gamma L_2|_{A^*} = 0, \quad (39)$$

$$L_2 + m_n^* F_\gamma^* L_1|_A = 0 \quad (40)$$

because $X(\alpha_k) = G_n(\alpha_k)^{-1}$ and hence $\hat{u}_\gamma(\alpha_k) = \frac{A_G A_{C_i}}{B_\gamma K_1 K_2 m_n} |_{\alpha_k}$. Finally, (33)–(34) reduce to

$$L_2(-a) + (E_\gamma(a) + 1)F_\gamma(a)m_n(a)L_1(-a) = 0 \quad (41)$$

$$L_1(-a) = 1. \quad (42)$$

In fact (34) reduces to $L_1(-a) = -\delta \frac{b_\gamma A_{C_i}^* m}{A_G} |_{s=-a}$, but replacing this equation by (42) does not effect the ratios y_0^*/x_0 and $\frac{y_0^* + z_0 x_0^* U}{x_0 + z_0 y_0 U}$, hence it does not effect the controller formulas.

Note that, by [10]

$$C = m_d G^{-1} N_0^{-1} \frac{\Theta}{1 - m_n G^{-1} \Theta} \text{ where } \Theta = W_0 + m_d Q_2. \quad (43)$$

On the other hand, by (18)

$$\begin{aligned} \Theta &= f_\gamma \left[m_1^* u_\gamma - m_d \frac{y_0^* + \frac{s-a}{s+a} x_0^* U}{x_0 + \frac{s-a}{s+a} y_0 U} \right] \\ &= m_n^* G^{-*} W_1^* W_1 - f_\gamma m_d \frac{y_0^* + \frac{s-a}{s+a} x_0^* U}{x_0 + \frac{s-a}{s+a} y_0 U}. \end{aligned} \quad (44)$$

By substituting (25), (26), (37), (44) into (43), and simplifying the resulting expression, we obtain

$$C_{\text{subopt}}(s) = E_\gamma(s) m_d(s) \frac{N_0^{-1}(s) F_\gamma(s) L_1(s)}{1 + m_n(s) F_\gamma(s) L_1(s)} \quad (45)$$

where $\gamma = \rho$, and

$$L_U(s) = \frac{L_2(s) + L_1(-s)U(s)}{L_1(s) + L_2(-s)U(s)}, \quad U \in \mathcal{B} \quad (46)$$

and L_1, L_2 are the unique nonzero polynomials of degree at most $n + l$ satisfying (38)–(42). The controller structure is the same for the optimal case, except the free parameter U drops

$$C_{\text{opt}}(s) = E_\gamma(s) m_d(s) \frac{N_0^{-1}(s) F_\gamma(s) L(s)}{1 + m_n(s) F_\gamma(s) L(s)} \quad (47)$$

where $L(s) = L_2(s)/L_1(s)$ and L_1, L_2 are nonzero polynomials of degree at most $n + l - 1$ satisfying (38)–(40).

V. CONCLUDING REMARKS

In this paper, a simple expression is obtained for the optimal and suboptimal \mathcal{H}^∞ controllers for a class of unstable distributed plants, with rational weighting functions. We have shown that \mathcal{H}^∞ controllers are in the form (45), (47), where m_d, m_n, N_0 are inner and outer parts of the plant $P = m_n N_0 / m_d$; E_γ is given in terms of the sensitivity weighting function W_1 ; F_γ is computed from a spectral factorization which depends on γ, W_1 and W_2 ; and polynomials L_1 and L_2 are computed from a finite set of linear equations. These linear equations can be written directly as interpolation conditions on L_1 and L_2 , that are expressed in terms of m_n and F_γ evaluated at the zeros of E_γ , and at the poles and zeros of m_d . The number of equations is $2(n_1 + l + 1)$ (in the optimal case it is $2(n_1 + l)$), which is equal to the number of unknown coefficients of L_1 and L_2 .

REFERENCES

- [1] V. M. Adamjan, D. Z. Arov, and M. G. Krein, "Analytic properties of Schmidt pairs for a Hankel operator and generalized Shur-Takagi problem," *Math. USSR Sbornik*, vol. 15, pp. 31–73, 1971.
- [2] R. F. Curtain, " H^∞ control for distributed parameter systems: A survey," in *Proc. 29th CDC*, Honolulu, Hawaii, Dec. 1990, pp. 22–26.
- [3] J. Doyle, K. Glover, P. P. Khargonekar, and B. Francis, "State space solutions to standard H^2 and H^∞ control problems," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 831–847, 1989.
- [4] D. Enns, H. Özbay, and A. Tannenbaum, "Abstract model and controller design for an unstable aircraft," *J. Guidance Contr. Dynamics*, vol. 15, pp. 498–508, 1992.
- [5] C. Foias, A. Tannenbaum, and G. Zames, "Some explicit formulae for the singular values of a certain Hankel operators with factorizable symbol," *SIAM J. Math. Analysis*, vol. 19, pp. 1081–1091, 1988.
- [6] B. Francis, *A Course in H^∞ Control Theory* (Lecture Notes in Control and Information Sciences), vol. 88. Berlin: Springer-Verlag, 1987.
- [7] J.-C. Juang and E. A. Jonckheere, "On computing the spectral radius of the Hankel plus Toeplitz operator," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 1053–1059, 1988.
- [8] H. Özbay, " \mathcal{H}^∞ optimal controller design for a class of distributed parameter systems," *Int. J. Contr.*, vol. 58, no. 4, pp. 739–782, Oct. 1993.
- [9] H. Özbay, M. C. Smith, and A. Tannenbaum, "Mixed sensitivity optimization for a class of unstable infinite dimensional systems," *Linear Algebra and Its Applications*, vol. 178, pp. 43–83, 1993.
- [10] —, "On the optimal two block H^∞ compensators for distributed unstable plants," in *Proc. Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 1865–1869.
- [11] M. C. Smith, "On stabilization and existence of coprime factorizations," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 1005–1007, 1989.
- [12] G. Zames, A. Tannenbaum, and C. Foias, "Optimal H^∞ interpolation: a new approach," in *Proc. 25th IEEE Conf. Decis. Contr.*, Athens, Greece, Dec. 1986, pp. 350–355.
- [13] G. Zames and S. K. Mitter, "A note on essential spectrum and norms of mixed Hankel-Toeplitz operators," *Syst. Contr. Lett.*, vol. 10, pp. 159–165, 1988.

Least Squares Type Algorithms for Identification in the Presence of Modeling Uncertainty

Er-Wei Bai and Krishan M. Nagpal

Abstract—The celebrated least squares and LMS (least-mean-squares) are system identification approaches that are easily implementable, need minimal *a priori* assumptions, and have very nice identification properties when the uncertainty in measurements is only due to noises and not due to unmodeled behavior of the system. When there is uncertainty present due to unmodeled part of the system as well, however, the performance of these algorithms can be poor. Here we propose a “modified” weighted least squares algorithm that is geared toward identification in the presence of both unmodeled dynamics and measurement disturbances. The algorithm uses very little *a priori* information and is easily implementable in a recursive fashion. Through an example we demonstrate the improved performance of the proposed approach. Motivated by a certain worst-case property of the LMS algorithm, an H_∞ estimation algorithm is also proposed for the same objective of identification in the presence of modeling uncertainty.

I. INTRODUCTION

The least squares and the LMS (least-mean-squares) algorithms are perhaps the most widely used identification algorithms [4], [7], [9], [11]. Their advantages lie in their simple recursive implementation and minimal use of *a priori* information. These algorithms perform very well in applications where the modeled behavior of the system closely approximates the behavior of the system to be identified. The performance of both least squares and LMS algorithms may become poor, however, if the assumed model structure is not rich enough to capture the dynamics of the system. Here we propose two algorithms that are motivated by the worst-case properties of the least squares and the normalized LMS approaches. While retaining most of the nice properties of the standard least squares (such as recursive implementation), these methods are also suitable for system identification in the presence of modeling inaccuracies.

Recently there has been a strong emphasis in the control community on “identification for control design.” The objective of system identification there is to obtain a good nominal plant together with some error bounds that are suitable for robust control design methodologies [2], [5], [6], [8], [10], [12], [13]. The approaches that have been proposed so far are, from a practical point of view, quite complicated and often result in nominal models of a very large order. There are other approaches to combine identification with control design [14], [16]. In the present approach we do not try to obtain nominal models with “hard” error bounds but instead sacrifice such rigid and precise objectives for optimality in a certain worst-case sense and ease of implementation. They do not rely much on the *a priori* information and especially the “modified” least squares algorithm is implementable in a recursive manner. For clarity of explanations, we will consider single input/single output systems. The approach is immediately generalizable to multi-input/multi-output systems.

The paper is organized as follows: in the next section we review some worst case properties of the standard least squares and the normalized LMS algorithms. Section III contains the precise problem formulation, Section IV the main results, and the Section V the proofs.

Manuscript received April 12, 1994.

The authors are with the Department of Electrical and Computer Engineering, The University of Iowa, Iowa City, IA 52243 USA.

IEEE Log Number 9408795.

II. REVIEW OF THE LEAST SQUARES AND LMS ALGORITHMS

Suppose we observe an output sequence y_i of a system with an ARMA structure

$$y_i = \phi_i' \theta + \xi_i, \quad i = 0, 1, \dots, N. \quad (2.1)$$

Here y_i is the output at the i th time instant, ξ_i is the measurement noise at the i th instant, $\phi_i \in R^n$ is a known regressor vector composed of past inputs and outputs, and $\theta \in R^n$ is the unknown vector to be identified that describes the system behavior. We will also use the following notation

$$Y := [y_0, y_1, \dots, y_N]', \quad \Phi := [\phi_0, \phi_1, \dots, \phi_N]' \\ \xi := [\xi_0, \xi_1, \dots, \xi_N]'. \quad (2.2)$$

Let \mathcal{A} be an identification algorithm. For the identifier \mathcal{A} , $\hat{\theta}(i)$ will be used to denote the estimate of θ obtained using measurements up to the stage i , i.e., $\hat{\theta}(i) = \mathcal{A}(y_0, y_1, \dots, y_i)$.

The next lemma describes some well-known properties of the least squares algorithm that are pertinent for our later analysis [1], [9].

Lemma 2.1 (Least Squares Properties): For given data Y and Φ , the least squares estimate of θ given by $\hat{\theta}(N) := (\Phi' \Phi)^{-1} \Phi' Y$ is optimal for the following two cost functions

$$\begin{aligned} \text{a)} \quad & \hat{\theta}(N) = \arg \inf_{\theta} \{ \|\xi\|^2 : Y = \Phi \theta + \xi \} \\ \text{b)} \quad & \hat{\theta}(N) = \arg \inf_{\theta} \sup_{\xi} \frac{\|\Phi \theta - Y\|^2}{\|\xi\|^2} \end{aligned}$$

The first part of the above lemma states that the least squares estimate of θ is that value of θ which explains the data with minimum amount (in l_2 norm sense) of noise. In the literature, this approach is sometimes referred to as total least squares [3]. Since in this case the uncertainty is only due to noises, the least squares estimate would be the answer to the question “what is the θ that explains the data with minimum amount of uncertainty?” In the next section we will ask this very question but also include the contribution from unmodeled behavior of the system in describing amount (norm) of uncertainty. To interpret the second part of the above lemma, consider the problem of estimating $\phi_i' \theta$ for $i \in [0, N]$ using the measurements y_0 to y_N . The cost function to be minimized is the worst case amplification in the l_2 norm from noises $\{\xi_i\}$ to the estimation errors $\phi_i'(\theta - \hat{\theta})$. This is a discrete-time H_∞ smoothing (noncausal estimation) problem since $\hat{\theta}$ is obtained using the entire measurement sequence. The second part of the above lemma states that if $\hat{\theta}(N)$ is the least squares estimate of θ , then $\phi_i' \hat{\theta}(N)$ is the best worst case (in H_∞ sense) smoother estimate of $\phi_i' \theta$. Thus the least squares estimate can be viewed as an optimal H_∞ smoother estimate.

Similar results can be said about the normalized LMS algorithm. For a given identifier \mathcal{A} , we will define the filtered error at stage i as $e_{i+1} = \phi_{i+1}'(\theta - \hat{\theta}(i))$.

Lemma 2.2 (Normalized LMS Property): For any given $\mu > 0$, consider the following normalized LMS algorithm

$$\hat{\theta}(i+1) := \hat{\theta}(i) + \frac{\mu \phi_{i+1}'}{1 + \mu \phi_{i+1}' \phi_{i+1}} (y_{i+1} - \phi_{i+1}' \hat{\theta}(i)), \quad \hat{\theta}(0) = \theta_0.$$

Then the above estimator achieves the following optimal worst-case cost

$$\inf_{\theta \in \mathcal{A}(Y)} \sup_{\xi} \frac{\|e\|^2}{\|\xi\|^2 + \frac{1}{\mu} \|\theta - \theta_0\|^2}.$$

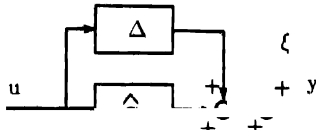


Fig. 1.

In the above, θ_0 represents the *a priori* estimate of θ . Thus the normalized LMS algorithm minimizes the worst-case l_2 gain from the “uncertainty” (in terms of *a priori* knowledge of θ and the disturbances) to the filtered error. The normalized LMS algorithm, however, is not the only algorithm that minimizes the above cost function [7]. The cost criterion in the above lemma and part b) of Lemma 2.1 are similar in the sense that in both cases one wants identifier for θ to give a good estimate for $\phi'_i \theta$. In part b) of Lemma 2.1, however, the estimate of θ used in predicting $\phi'_i \theta$ is based on the entire sequence of measurements available while in Lemma 2.2 the estimate of θ used in predicting $\phi'_i \theta$ is the one based on only the measurements up to time instant i .

III. THE PROBLEM FORMULATION

In this section we will formulate two approaches for identification in presence of both the measurement noises as well as unmodeled dynamics. It is worth reemphasizing that the objectives of the present approach are i) to make as few “hard” assumptions as possible (such as a bound on the noise, or an H_∞ bound on the unmodeled behavior), and ii) to come up with an algorithm that is easily implementable.

We will assume an additive uncertainty model (see Fig. 1) for describing the input-output behavior of the system

$$y = \hat{G}u + \Delta u + \xi. \quad (3.3)$$

Here \hat{G} is the nominal model that we wish to identify, ξ is the measurement noise, and Δ is an unknown stable linear time invariant system that captures the unmodeled dynamics of the system. Assuming an ARMA structure for the modeled part \hat{G} for the structure, one can write the above equation as

$$y_i = \phi'_i \theta + (\Delta u)_i + \xi_i, \quad i = 0, 1, \dots, N \quad (3.4)$$

where ϕ_i , the regressor vector is composed of known inputs and outputs and θ is the unknown parameter we wish to identify. Here we are also assuming that only a finite number of observations (N) are available for identification. Let the (unknown) impulse response of the uncertainty Δ be $\delta_0, \delta_1, \dots$. We can then rewrite (3.4) more compactly as

$$Y = \Phi \theta + U \delta + \xi \quad (3.5)$$

where Y , Φ , and ξ are defined in (2.2), and

$$\delta := \begin{bmatrix} \delta_0 \\ \delta_1 \end{bmatrix}, \quad U = \begin{bmatrix} u_0 & 0 \\ u_1 & u_0 \end{bmatrix} \quad (3.6)$$

Motivated by part a) of Lemma 2.1, we now propose our first approach for identification of θ .

Problem 1—Modified Weighted Least Squares Problem (MWLSP): Given are two positive definite weighting matrices $W_\delta = W_\delta$ and $W_\xi = W_\xi$. Find θ so as to

$$\begin{aligned} &\text{minimize} \quad \delta^T W_\delta \delta + \xi^T W_\xi \xi \\ &\text{subject to} \quad Y = \Phi \theta + U \delta + \xi. \end{aligned}$$

Before discussing the role of W_δ and W_ξ in the above problem, we will briefly discuss the underlying motivation behind the above

problem formulation. The motivation behind the above problem is qualitatively very similar to the property a) of the standard least squares stated in the Lemma 2.1. Let

$$J(W_\delta, W_\xi) := \|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2.$$

The quantity $J(W_\delta, W_\xi)$ can be interpreted as the size (or norm) of uncertainty (due to both unmodeled behavior and the measurement noises). Thus in the modified weighted least squares problem (MWLSP) we want to “find θ that explains the data with minimum amount of uncertainty.” In the standard least squares one asks this very question but there all the uncertainty is attributed to noises only.

The weighting matrices W_δ and W_ξ are chosen by designers to measure the individual contributions of the unmodeled behavior and the disturbances. This can be used to incorporate some prior knowledge. We discuss some cases here.

Case 1) Consider the following choice of weights

$$W_\delta = \lambda I, \quad \text{and} \quad W_\xi = (1 - \lambda)I, \quad 0 < \lambda < 1$$

where I is the $(N+1)$ -dimensional identity matrix. The quantity λ is chosen to reflect the relative contribution of unmodeled behavior and the disturbances. If $\lambda \rightarrow 0$ then one expects that most of the uncertainty is due to δ or the unmodeled behavior while $\lambda \rightarrow 1$ corresponds to the belief that most of the uncertainty is due to noises and not due to the unmodeled behavior. Indeed, as can be easily observed, the above method reduces to the standard least squares when $W_\delta = \lambda I \rightarrow I$ and $W_\xi = (1 - \lambda)I \rightarrow 0$.

Case 2) The weightings W_δ can also be used to shape the accuracy requirements of identification with respect frequency. For example, consider a situation where it is desired to have a more accurate description of the system behavior at low frequencies. In this case, we would want the unmodeled dynamics (system Δ in Fig. 1), which is “consistent” with the identified model in explaining the data, to be small at low frequencies. Let $\{h_i\}$ be the impulse response of a low pass filter. Consider now the following choice for W_δ

$$W_\delta = \begin{bmatrix} h_0 & 0 & & 0 & 0 \\ h_1 & h_0 & & 1 & h_0 \\ & & \ddots & & \\ h_N & & & h_0 & h_{N-1} & h_0 \end{bmatrix}$$

With the above choice of W_δ , the norm of the unmodeled uncertainty ($\|\delta\|_{W_\delta}^2 = \delta^T W_\delta \delta$) is given more weighting in the low frequency region.

Case 3) The unknown system is assumed to be stable. Then we would probably also desire the nominal model obtained from identification to be stable, and thus so would be the unmodeled dynamics Δ . Consequently, the impulse response $\{\delta_k\}$ of Δ converges to zero as $k \rightarrow \infty$. This *a priori* knowledge that $\delta_k \rightarrow 0$ as $k \rightarrow \infty$ can be incorporated in the identification process by suitable selection of W_δ and W_ξ . We next describe one such choice for W_δ and W_ξ . The weightings W_δ and W_ξ should reflect the fact that the “tail” contribution of $\{\delta_k\}$ for large k is much smaller than that of the disturbance sequence $\{\xi_k\}$. This can be easily incorporated into the weighting matrices by choosing them, for example, as follows

$$W_\delta = \text{diag}(\alpha, \alpha^2, \dots, \alpha^N) \quad \text{and} \quad W_\xi = cI$$

where c is some positive real number, I the identity matrix of appropriate size and $\alpha > 1$. With these choices of weightings, a unit ball of uncertainty defined as $B = \{(\delta, \xi): \|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2 \leq 1\}$ would have δ_k comparatively smaller than ξ_k for large k .

We shall observe in the next section that the optimal θ for the Problem 1 is also optimal in a worst-case sense, a property very similar to property b) of Lemma 2.1. To be more precise, the optimal θ for the MWLSP also satisfies

$$\text{optimal } \theta \text{ for MWLSP} = \arg \inf_{\theta \in \mathcal{A}(Y)} \sup_{\xi, \delta} \frac{\|\Phi\theta - \Phi\hat{\theta}\|^2}{\|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2}. \quad (3.7)$$

Property (3.7) has an obvious worst-case interpretation. Its usefulness can be seen from another angle as well. From a good model one would expect to be able to predict the uncorrupted system output accurately. From (3.4) it follows that for a good identification scheme, $\phi_i' \hat{\theta}$ would be close to $y_i - \xi_i$ for all i . Thus one meaningful criterion for system identification would be to find an identification algorithm that minimizes the worst-case gain from "uncertainty" to the error in predicting the actual system output or to find $\hat{\theta} = \mathcal{A}(Y)$ that achieves

$$J = \inf_{\hat{\theta} \in \mathcal{A}(Y)} \sup_{\xi, \delta} \frac{\|(\Phi\hat{\theta} + U\delta) - \Phi\hat{\theta}\|}{\{\|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2\}^{\frac{1}{2}}}. \quad (3.8)$$

It turns out that the solution to the above problem is not easily implementable. As we show next, however, the MWLSP algorithm minimizes an upper bound for J . To see this note that

$$\sup_{\xi, \delta} \frac{\|\Phi\hat{\theta} - (\Phi\theta + U\delta)\|}{\{\|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2\}^{\frac{1}{2}}} \leq \sup_{\xi, \delta} \frac{\|\Phi\hat{\theta} - \Phi\theta\|}{\{\|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2\}^{\frac{1}{2}}} + \sigma_{\max}(M_\delta^{-1}U'U^{-1}M_\delta^{-1})$$

where M_δ is the square root of W_δ , i.e., $M_\delta' M_\delta = W_\delta$ and σ_{\max} denotes the largest singular value. The second term in the above inequality is independent of the algorithm, while the first one is precisely the quantity that is minimized by MWLSP [see (3.7)]. Thus the MWLSP has an additional property of minimizing an upper bound for J , the worst case gain from "uncertainty" to the error in estimating the uncorrupted system output (output when there is no measurement noise).

Next we present the second problem we will consider. The motivation for the next problem comes from Lemma 2.2, a framework that follows from an immediate generalization of the worst-case property of the normalized LMS algorithm. Recall that for any given identifier \mathcal{A} , the filtered error at stage i is defined as $e_{i,i} = \phi_i'(\theta - \hat{\theta}(i))$.

Problem 2— H_∞ Filtering Identification Problem (HFIP): Given are two positive definite matrices W_δ and W_ξ and real numbers $\mu > 0$. Find the identification algorithm \mathcal{A} so that $\hat{\theta} = \mathcal{A}(Y)$ achieves

$$\inf_{\theta \in \mathcal{A}(Y)} \sup_{\xi, \delta, \theta} \frac{\|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2 + \frac{1}{\mu} \|\theta - \theta_0\|^2}{\|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2}$$

where the supremum over ξ, δ and θ is taken over all ξ, δ, θ that satisfy $Y = \Phi\theta + U\delta + \xi$.

Again the motivation is similar to before—we want to minimize the worst case amplification from "uncertainty" to the prediction error of the modeled part of the system. The previous discussions about the roles of W_δ and W_ξ also apply for this case.

IV. MAIN RESULTS

In this section, we present the solution to MWLSP and HFIP problems.

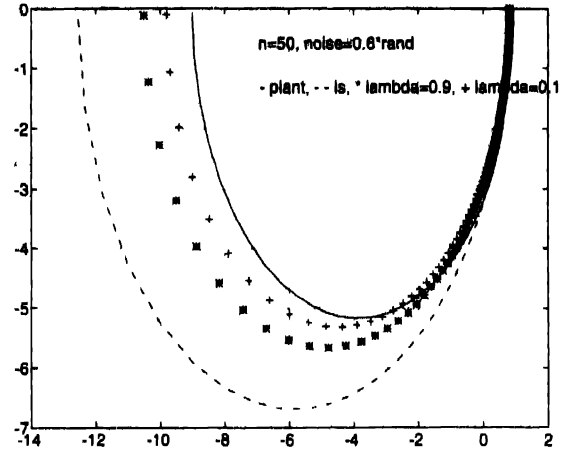


Fig. 2.

Theorem 4.1 (Solution of the MWLSP Problem). Consider the identification problem for a system whose input–output behavior is described by (3.5). For given input–output data Y , Φ and the weighting matrices W_δ and W_ξ , if Φ is full column rank, then

1)

$$\hat{\theta}_{mwls} = (\Phi'Q\Phi)^{-1}\Phi'QY \quad (4.9)$$

with

$$Q = (W_\xi^{-1} + U'W_\delta^{-1}U)^{-1} \quad (4.10)$$

solves the MWLSP problem. Moreover, the solution is unique and also admits a recursively implementable algorithm as shown in Lemma 4.2.

2) $\hat{\theta}_{mwls}$ also achieves the following worst-case cost

$$\inf_{\theta \in \mathcal{A}(Y)} \sup_{\xi, \delta} \frac{\|\Phi\theta - \Phi\hat{\theta}\|^2}{\|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2}.$$

The solution $\hat{\theta}_{mwls}$ to the MWLSP is basically a weighted least squares estimate of θ with the weight Q . Unlike the standard weighted least squares problems, however, here the weighting matrix Q also depends on the input. Being a weighted least squares algorithm, it possesses all the nice properties of the least squares, such as recursive implementation as is discussed in the Lemma 4.2.

Before presenting the next results, we present an example to demonstrate the performance of the MWLSP algorithm. The system to be identified is the following second-order system

$$y_i = -1.3y_{i-1} - 0.4y_{i-2} + 1.5u_{i-1} + 0.6u_{i-2} + \xi_i.$$

For simulation, ξ_i was assumed to be an independent uniformly distributed random variable in $[0, 0.6]$, input $u_i = \sin(i) + 2\cos(2.3 * i)$ and $N = 50$. The nominal model is assumed to be of the first order. Simulations were done for two choices of weights: in one case the weights were chosen as $W_\delta = 0.9I$, $W_\xi = 0.1I$ (from the discussion of the Case 1) in the previous section, this choice corresponds to an *a priori* assumption that measurement noises are more important/dominant than unmodeled dynamics) and in another case chosen as $W_\delta = 0.1I$, $W_\xi = 0.9I$ (similarly corresponds to an *a priori* belief that unmodeled dynamics are more important than noises). Fig. 2 shows the Nyquist plots of the true but unknown system (solid), the least squares estimate (dashdot) and the MWLSP estimates proposed in the paper with $W_\delta = 0.9I$, $W_\xi = 0.1I$ (**) and $W_\delta = 0.1I$, $W_\xi = 0.9I$ (+), respectively. As one might have expected, for both these choices of weights, the proposed algorithm outperforms the least squares algorithm.

The calculation of $\hat{\theta}_{mwls}$ in Theorem 4.1 can be made recursive. Before summarizing the recursive procedure, we describe some

notation used in the recursion. As defined previously, W_δ and W_ξ are $(N+1) \times (N+1)$ positive definite matrices where N denotes the number of data points. Let $W_\delta(k)$ and $W_\xi(k)$ denote the upper left $k \times k$ submatrix of W_δ and W_ξ respectively, i.e.,

$$W_\xi = \begin{bmatrix} W_\xi(k) & * \\ * & W_\delta = W_\delta(k) & * \end{bmatrix}$$

where $*$ indicates the remaining portions of the respective matrices. Because of the symmetry of the weighting matrices $W_\delta(k)$ and $W_\xi(k)$, their inverses can be written as

$$W_\xi^{-1}(k) := \begin{bmatrix} A(k-1) & a_k \\ a_k' & a_{k,k} \end{bmatrix}$$

$$W_\delta^{-1}(k) := M_\delta^{-1}(k) V_\delta^{-1}(k)$$

with $M_\delta^{-1}(k) := \begin{bmatrix} B(k-1) & 0 \\ b_k' & b_{k,k} \end{bmatrix}$ being the Cholesky factor of $W_\delta^{-1}(k)$, where $a_{k,k}$ and $b_{k,k}$ are scalars, and a_k and b_k are column vectors. The data is given by

$$Y(k) := \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{bmatrix}, \quad \Phi(k) := \begin{bmatrix} \phi_0' \\ \phi_1' \\ \vdots \\ \phi_k' \end{bmatrix}$$

$$\begin{bmatrix} u_0 & 0 & \cdots & 0 \\ u_1 & u_0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ u_k & u_{k-1} & \cdots & u_0 \end{bmatrix} \begin{bmatrix} C'(k-1) & 0 \\ c_k' & c_{k,k} \end{bmatrix}$$

with $c_{k,k}$ being a scalar and c_k a column vector. Define

$$Z(k) := U(k) V_\delta^{-1}(k) = \begin{bmatrix} C'(k-1)B(k-1) & 0 \\ c_k' B(k-1) + u_0 b_k' & u_0 b_{k,k} \end{bmatrix}$$

$$= \begin{bmatrix} Z(k-1) & 0 \\ z_k' & z_{k,k} \end{bmatrix}$$

$$Q(k) := (W_\xi^{-1}(k) + U(k) W_\delta^{-1}(k) U'(k))^{-1}$$

$$= (A(k) + Z(k) Z'(k))^{-1} = L'(k) L(k)$$

with

$$L^{-1}(k) := \begin{bmatrix} L^{-1}(k-1) & 0 \\ l_k' & l_{k,k} \end{bmatrix}$$

or

$$L(k) = \begin{bmatrix} L(k-1) \\ \frac{1}{l_{k,k}} l_k' L(k-1) \end{bmatrix}$$

The next lemma summarizes the recursive formula for obtaining θ_{mwt} for the MWLSP stated in the Theorem 4.1.

Lemma 4.2 (Recursive Algorithm for MWLSP): A recursive algorithm for obtaining

$$\hat{\theta}_{\text{mwt}}(k) = (\Phi'(k) Q(k) \Phi(k))^{-1} \Phi'(k) Q(k) Y(k)$$

is given as follows.

Given initial data y_0, ϕ_0' , calculate

$$L(0) = (A(0) + Z(0)Z'(0)B(0)B'(0))^{1/2}, Y(0) = L(0)y_0,$$

$$\bar{\Phi}(0) = L(0)\phi_0'.$$

Step 1) At k th step, for any given new data u_k, y_k , and ϕ_k' , calculate

$$Z(k) = U(k)B(k), \quad l_k = L(k-1)(a_k + Z(k-1)z_k)$$

$$l_{k,k} = \sqrt{a_{k,k} + z_{k,k}^2 + z_k' z_k - l_k' l_k}$$

$$\bar{y}_k = \frac{1}{l_{k,k}}(y_k - l_k' \bar{Y}(k-1)),$$

$$\bar{\phi}_k' = \frac{1}{l_{k,k}}(\phi_k' - l_k' \bar{\Phi}(k-1)).$$

Step 2)

$$\hat{\theta}_{\text{mwt}}(k) = \theta_{\text{mwt}}(k-1)$$

$$+ \frac{P(k-1)\bar{\phi}_k}{1 + \bar{\phi}_k' P(k-1)\bar{\phi}_k} (\bar{y}_k - \bar{\phi}_k' \hat{\theta}_{\text{mwt}}(k-1))$$

$$P(k) = P(k-1) \frac{P(k-1)\bar{\phi}_k \bar{\phi}_k' P(k-1)}{1 + \bar{\phi}_k' P(k-1)\bar{\phi}_k}$$

Step 3) Set

$$Y(k) = \begin{bmatrix} \bar{Y}(k-1) \\ \bar{y}_k \end{bmatrix}, \quad \bar{\Phi}(k) = \begin{bmatrix} \bar{\Phi}(k-1) \\ \bar{\phi}_k' \end{bmatrix}$$

$$L(k-1)$$

$$L(k) = \frac{1}{l_{k,k}} l_k' L(k-1)$$

Step 4) Set $k+1 = k$ and go to Step 1.

Note that $l_{k,k}$ is away from zero since $W_\xi(k) > 0$ for all k .

We next present the solution to the H_∞ filtering identification problem. The following result can be obtained by immediate application of the standard results for the discrete-time H_∞ filtering problem (see, for example, [7], [15]). Recall that for any given identifier \mathcal{A} , the filtered error at stage i is defined as $e_{f,i} = \phi_i'(\theta - \theta(i))$.

Theorem 4.3 (Solution to the HFIP): Given are a positive definite matrix W_δ and a real number $\mu > 0$. There exists an identification algorithm \mathcal{A} so that $\hat{\theta} = \mathcal{A}(Y)$ achieves

$$J := \sup_{\theta} \left(\|\hat{\theta} - \theta\|_{W_\delta}^2 + \|\hat{\theta} - \theta\|^2 + \frac{1}{\mu} \|\theta - \hat{\theta}\| \right)$$

$$+ h_i' h_i - \frac{1}{\mu} l_i' l_i > 0, \quad i = 1, \dots, N \quad (4.11)$$

where

$$P(0) = \begin{bmatrix} \mu I & 0 \\ 0 & W_\delta^{-1} \end{bmatrix}, \quad h_i = [\phi_i' u_i \ u_{i-1} \ \cdots \ u_0 \ 0]$$

$$l_i = [\phi_i' \ 0 \ \cdots \ 0]$$

and $P(i)$ satisfy the following recursion

$$P^{-1}(i+1) = P^{-1}(i) + h_i' h_i - \frac{1}{\mu} l_i' l_i. \quad (4.12)$$

Moreover, if (4.11) is satisfied, the following identifier achieves $J < \gamma^2$

$$\begin{bmatrix} \hat{\theta}(i+1) \\ \hat{\delta}(i+1) \end{bmatrix} = \begin{bmatrix} \hat{\theta}(i) \\ \hat{\delta}(i) \end{bmatrix} + \frac{P(i+1)h_{i+1}}{1 + h_{i+1}' P(i+1)h_{i+1}} \left\{ y_{i+1} - h_{i+1}' \begin{bmatrix} \hat{\theta}(i) \\ \hat{\delta}(i) \end{bmatrix} \right\}$$

with $[\theta(0), \delta(0)] = [\theta_0, 0]$.

From an implementation point of view, the above algorithm is not as nice as the MWLSP algorithm for two reasons—i) the order of the estimator is equal to the size of the unknown vector θ plus the number of data points (while in recursive implementation of MWLSP, the order of estimator is just equal to the size of the unknown vector θ) and ii) to check whether (4.11) holds, one has to iterate with respect to γ . The recursive computation of $P(i)$ in (4.12) can be simplified, however, using the Matrix Inversion Lemma so that in obtaining $P^{-1}(i+1)$ from $P^{-1}(i)$, one only needs to invert a 2×2 matrix.

V. PROOFS

In this section, we will give a brief outline of the proofs for the Theorem 4.1 and the Lemma 4.2.

In the proofs, we will use an observation that we present in the next lemma. Define a set

$$\mathcal{C} = \{\mathcal{A}: \mathcal{A}(Y) = \theta \text{ if } \delta = 0 \text{ and } \xi = 0\}. \quad (5.13)$$

The set \mathcal{C} is referred to as the set of correct algorithms which represents all the identification algorithms that give the exact estimate θ if both δ and ξ are absent. The following lemma is quite straightforward.

Lemma 5.1: Let \mathcal{A} be any identification algorithm $\mathcal{A}(Y) = \theta$. Then, $\sup_{\xi, \delta} \frac{\|\Phi\hat{\theta} - \Phi\theta\|^2}{\|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2}$ is bounded only if $\mathcal{A} \in \mathcal{C}$.

Proof of Theorem 4.1:

Part 1): From (3.5), ξ can be written in terms of θ and δ as

$$\xi = Y - \Phi\theta - U\delta.$$

Thus

$$\begin{aligned} J_1 &:= \|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2 \\ &= \delta' W_\delta \delta + (Y - \Phi\theta - U\delta)' W_\xi (Y - \Phi\theta - U\delta). \end{aligned}$$

Taking derivative of J_1 with respect to θ and δ , respectively, and setting them to zero yield

$$\begin{aligned} 0 &= 2W_\delta \delta - 2U' W_\xi (Y - \Phi\theta - U\delta), \\ 0 &= -2\Phi' W_\xi (Y - \Phi\theta - U\delta). \end{aligned}$$

Solving these two equations for θ , we have

$$\hat{\theta}_{\text{mwls}} = (\Phi' Q \Phi)^{-1} \Phi' Q Y$$

with

$$\begin{aligned} Q &= W_\xi - W_\xi U (W_\delta + U' W_\xi U)^{-1} U' W_\xi \\ &= W_\xi (I + U' W_\delta^{-1} U' W_\xi)^{-1} \\ &= (W_\xi^{-1} + U' W_\delta^{-1} U)^{-1}. \end{aligned} \quad (5.14)$$

The uniqueness follows from convexity. This completes the proof.

Part 2) Let

$$\Phi = U_1 \begin{bmatrix} D_1 & | & 0 \end{bmatrix} U_1' \quad (5.15)$$

be the singular value decomposition of Φ , where U_1 and V_1 are unitary matrices and D_1 is a diagonal positive definite matrix. Let \bar{Q} be defined as

$$Q = U_1 \bar{Q} U_1' = U_1 \begin{bmatrix} \bar{Q}_{11} & \bar{Q}_{12} \\ \bar{Q}_{12}' & \bar{Q}_{22} \end{bmatrix} U_1'. \quad (5.16)$$

Also, M_δ and M_ξ are defined as $M_\delta' M_\delta = W_\delta$ and $M_\xi' M_\xi = W_\xi$. Now, the proof is divided into two parts. In the first part, we will show that if $\hat{\theta} = \mathcal{A}(Y)$ and $\mathcal{A} \in \mathcal{C}$, then

$$\sup_{\xi, \delta} \|\Phi\hat{\theta} - \Phi\theta\| \begin{bmatrix} M_\delta \delta \\ M_\xi \xi \end{bmatrix} \geq \frac{1}{\sigma_{\min}(\bar{Q}_{11})}. \quad (5.17)$$

In the second part, we will show that for the MWLSP estimate $\hat{\theta}_{\text{mwls}}$,

$$\sup_{\xi, \delta} \|\Phi\hat{\theta}_{\text{mwls}} - \Phi\theta\| \begin{bmatrix} M_\delta \delta \\ M_\xi \xi \end{bmatrix} \leq \frac{1}{\sigma_{\min}(\bar{Q}_{11})}. \quad (5.18)$$

We now show (5.17). Let θ be the unknown vector to be identified and let θ^* be any vector of the same dimension. Let δ and ξ be as follows

$$\begin{bmatrix} \delta \\ \xi \end{bmatrix} = \begin{bmatrix} W_\delta^{-1} & 0 \\ 0 & W_\xi^{-1} \end{bmatrix} \begin{bmatrix} U_1' \\ I \end{bmatrix} Q \Phi (\theta^* - \theta).$$

For the above δ and ξ , from (3.5) and (5.14), $Y = \Phi\theta + U\delta + \xi = \Phi\theta^*$. The resulting data Y and Φ are as if θ^* were the true value

of the unknown parameter and there were no uncertainties/noises present. From Lemma 5.1, thus for any $\mathcal{A} \in \mathcal{C}$ (the class of "correct algorithms"), the estimate $\hat{\theta} = \mathcal{A}(Y) = \theta^*$. Since from Lemma 5.1, our search is restricted to the class of "correct algorithms," we obtain using (5.14)–(5.16) that

$$\begin{aligned} \|\Phi\mathcal{A}(Y) - \Phi\theta\| & \begin{bmatrix} M_\delta \delta \\ M_\xi \xi \end{bmatrix} \\ &= \begin{bmatrix} (\theta^* - \theta)' U_1' D_1 D_1 U_1' (\theta^* - \theta) \\ (\theta^* - \theta)' U_1' \bar{Q}_{11} D_1 U_1' (\theta^* - \theta) \end{bmatrix} = \begin{bmatrix} \zeta' \zeta \\ \zeta' \bar{Q}_{11} \zeta \end{bmatrix} \end{aligned}$$

where $\zeta = D_1 U_1' (\theta^* - \theta)$. Since $D_1 U_1'$ is nonsingular and θ^* is arbitrary, ζ is arbitrary. Thus

$$\sup_{\xi, \delta} \|\Phi\hat{\theta} - \Phi\theta\| \begin{bmatrix} M_\delta \delta \\ M_\xi \xi \end{bmatrix} > \frac{1}{\sigma_{\min}(\bar{Q}_{11})}$$

or that (5.17) holds. We now show (5.18). Note that

$$\theta_{\text{mwls}} = (\Phi' Q \Phi)^{-1} \Phi' Q Y = \theta + (\Phi' Q \Phi)^{-1} \Phi' Q (U\delta + \xi),$$

$$\text{and } \Phi(\Phi' Q \Phi)^{-1} \Phi' = U_1 \begin{bmatrix} \bar{Q}_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} U_1'.$$

Hence

$$\|\Phi\theta_{\text{mwls}} - \Phi\theta\|^2 = \left\| S \begin{bmatrix} M_\delta \delta \\ M_\xi \xi \end{bmatrix} \right\|^2$$

where $S = U_1 \begin{bmatrix} \bar{Q}_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} U_1' Q (U' M_\delta^{-1} U' M_\xi^{-1})$. Thus $S S' = U_1 \begin{bmatrix} \bar{Q}_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} U_1'$ and from above it follows that

$$\sup_{\xi, \delta} \frac{\|\Phi\hat{\theta}_{\text{mwls}} - \Phi\theta\|^2}{\|\delta\|_{W_\delta}^2 + \|\xi\|_{W_\xi}^2} \leq \sigma_{\max}(\bar{Q}_{11}^{-1}) = \frac{1}{\sigma_{\min}(\bar{Q}_{11})}.$$

This completes the proof. \square

Proof of Lemma 4.2. Note that

$$\theta_{\text{mwls}}(k) = (\Phi'(k) Q(k) \Phi(k))^{-1} \Phi'(k) Q(k) Y(k)$$

where $Q(k)$ is symmetric and positive definite. Let $Q(k) = L'(k) L(k)$, we have

$$\begin{aligned} \theta_{\text{mwls}}(k) &= \arg \min_{\theta} \|L(k)(Y(k) - \Phi(k)\theta)\| \\ &= \arg \min_{\theta} \|\bar{Y}(k) - \Phi(k)\theta\| \end{aligned}$$

where $\bar{Y}(k) = L(k)Y(k)$ and $\Phi(k) = L(k)\Phi(k)$. From above, we see that $\hat{\theta}_{\text{mwls}}(k)$ can be obtained recursively (using the standard recursive least squares algorithm) if $\bar{Y}(k)$ and $\Phi(k)$ can be calculated so. It can be easily verified that the algorithm shown in Lemma 4.2 recursively estimates $Y(k)$ and $\Phi(k)$ and then generates the estimate by implementing the recursive least squares approach on the new data. \square

REFERENCES

- [1] E. W. Bai and R. Tempo, "Least squares parameter estimator," in *Proc. Amer. Contr. Conf.*, 1993.
- [2] J. Chen, C. N. Nett, and M. Fan, "Worst-case system identification in H_∞ validation of *a priori* information, essentially optimal algorithms, and error bounds," in *Proc. Amer. Contr. Conf.*, 1992, pp. 251–257.
- [3] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1984.
- [4] G. Goodwin and K. Sin, *Adaptive Filtering, Prediction and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
- [5] G. Gu and P. Khargonekar, "Linear and nonlinear algorithms for identification in H_∞ with error bounds," *IEEE Trans. Automat. Contr.*, vol. 39, 1992.

- [6] A. J. Helmicki, C. A. Jacobson, and C. N. Nett, "Control oriented system identification: a worst case/deterministic approach in H_∞ ," *IEEE Trans Automat. Contr.*, vol. 36, pp. 1163-1176, 1991.
- [7] B. Hassibi, A. H. Sayed, and T. Kailath, "LMS is H_∞ optimal," in *Proc 32nd Conf. Decis. Contr.*, San Antonio, TX, 1993.
- [8] B. Kacewicz and M. Milanese, "Optimal finite sample experiment design in worst case l_1 system identification," in *Proc IEEE Conf. Decis. Contr.*, 1992, pp. 56-61.
- [9] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
- [10] P. M. Makila and J. R. Partington, "Worst-case identification from closed-loop time series," in *Proc Amer. Contr. Conf.*, 1992, pp. 301-306.
- [11] J. P. Norton, "Identification and application of bounded parameter models," *Automatica*, vol. 23, pp. 497-597, 1987.
- [12] K. Poola and A. Tikku, "On the time complexity of worst-case system identification," preprint, 1992.
- [13] D. Tse, M. Dahleh, and J. Tsitsiklis, "Optimal asymptotic identification under bounded disturbance," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 1176-1190, 1991.
- [14] P. M. J. Van den Hof, R. J. P. Schrama, O. H. Bosgra, and R. A. de Callafon, "Identification of normalized coprime plant factors for iterative model and controller enhancement," in *Proc. Conf. Decis. Contr.*, 1993, pp. 2839-2844.
- [15] I. Yaesh and U. Shaked, "A transfer function approach to the problem of discrete time systems. H_∞ -optimal linear control and filtering," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1264-1271, 1991.
- [16] Z. Zhang, R. Bitmead, and M. Gevers, "Disturbance rejection on-line refinement of controllers by closed loop modeling," in *Proc. Amer. Contr. Conf.*, 1992, pp. 2829-2833.

Square-Root Bryson-Frazier Smoothing Algorithms

PooGyeon Park and Thomas Kailath

Abstract—Some new square-root algorithms for Bryson-Frazier smoothing formulas are suggested: square-root algorithms and a fast square-root (or so-called Chandrasekhar type) algorithm. The new square-root algorithms use square-root arrays composed of smoothed estimates and their error covariances. These algorithms provide many advantages over the conventional algorithms with respect to systolic array and parallel implementations as well as numerical stability and conditioning. For the case of constant-parameter systems, a fast square-root algorithm is suggested, which requires less computation than others.

I. INTRODUCTION

Square-root (or factorized, as they are sometimes called) algorithms for state-space estimation have been found to have several advantages over the conventional equation-based algorithms in numerical stability, conditioning, and amenability to parallel and systolic implementation. While such algorithms for prediction and filtering have by now been studied quite extensively (see, e.g., [1]-[8]), the picture is not quite as complete for smoothed estimates. Square-root smoothing algorithms in the literature have various advantages and disadvantages (see, e.g., [4], [9]-[15]); all of them require certain inversion on back-

substitution steps, and none of them are particularly well suited for parallel implementation.

In another direction, there has been considerable effort in speeding up Kalman filtering [7], [16]. While the conventional equation-based or square-root algorithms require $O(n^2)$ flops per iteration, fast square-root (or Chandrasekhar type) algorithms for time-invariant systems [17]-[20] require only $O(n(p+q))$ flops, where p and q are the number of outputs and the displacement rank of Riccati solutions $P_{i|i-1}$, (i.e., $\text{rank}\{P_{i+1|i} - P_{i|i-1}\}$), respectively. For smoothing, however, there seems to be no fast square-root algorithms in the literature.

In our paper, we shall present some new square-root smoothing algorithms which are more suitable for parallel implementation than any algorithm in the literature and also a fast square-root smoothing algorithm for constant-parameter systems. As a target, we shall consider the basic Bryson-Frazier smoothing formulas [21].

Section II reviews Kalman filtering and some recast versions [22], and Section III describes the Bryson-Frazier smoothing formulas, presents several square-root versions, and briefly compares them on the basis of constraints, memory size, and computational speed.

For convenience we first introduce some notational conventions.

Square-Root Factors: Given a positive definite matrix A , $A > 0$, a square-root factor will be defined as any matrix, say $A^{1/2}$, such that $A = (A^{1/2})(A^{1/2})^*$, where the "*" denotes matrix transpose. Such square-root factors are clearly not unique. In most applications, the triangular form is preferred. For convenience we shall also write $(A^{1/2})^* = A^{*/2}$, $(A^{1/2})^{-1} = A^{-1/2}$, $(A^{-1/2})^* = A^{-*/2}$.

Square-Root Forms: Given a matrix A , assume that we apply a unitary operator Θ to the A so as to get some special form of a matrix B such as $A\Theta = B$, where $AA^* = BB^*$, then we shall call the A a pre-array and the B a post-array.

Matrix Inversion Lemma $(A + BC^*D)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$.

State-Space Model: $x_{i+1} = F_i x_i + G_i u_i$, $y_i = H_i x_i + v_i$, $i \geq 0$, where $F_i \in \mathbb{C}^{n \times n}$, $H_i \in \mathbb{C}^{p \times n}$, $G_i \in \mathbb{C}^{n \times m}$, and $\{x_0, u_i, v_i\}$ are independent zero-mean white Gaussian random variables with the covariances $\{\Pi_0, Q_i, R_i\}$. The matrices $\{F_i, G_i, H_i, Q_i, R_i, \Pi_0\}$ are assumed to be known. For simplicity, we shall define $\hat{x}_{i|i} \triangleq$ the linear least-squares estimate of x_i given $\{y_0, \dots, y_i\}$, $P_{i|i} \triangleq E[x_i - \hat{x}_{i|i}][x_i - \hat{x}_{i|i}]^*$, the error covariance of the estimate $\hat{x}_{i|i}$. For compactness we shall write, unless necessary for emphasis or comparison, $P_{i|i-1} \triangleq P_i$.

II. SQUARE-ROOT KALMAN FILTERING

Given the standard state-space model described above, we have the following recursions: for measurement updates

$$\hat{x}_{i|i} = \hat{x}_i + K_i \epsilon_i, \quad P_{i|i} = P_i - K_i R_i^{-1} K_i^*$$

and for time updates

$$x_{i+1} = F_i \hat{x}_{i|i}, \quad P_{i+1} = F_i P_{i|i} F_i^*$$

where $K_i \triangleq P_i H_i^* R_i^{-1}$, $R_{i-1} \triangleq R_i + H_i P_i H_i^*$, $\epsilon_i \triangleq y_i - H_i \hat{x}_i$, and the P_i are propagated via the Riccati difference equation: $P_{i+1} = F_i P_i F_i^* + G_i Q_i G_i^* - F_i K_{i-1} R_{i-1}^{-1} K_{i-1}^* F_i^*$, $P_0 = \Pi_0$.

When Π_0 is very large, it is often preferred to propagate P_i^{-1} , which can be performed when $\Pi_0 > 0$, $R_i > 0$, and the F_i are invertible. The resulting so-called "Information Filter" formulas are as follows: for measurement updates

$$P_{i|i}^{-1} \hat{x}_{i|i} = P_i^{-1} \hat{x}_i + H_i^* R_i^{-1} y_i, \quad P_{i|i}^{-1} = P_i^{-1} + H_i^* R_i^{-1} H_i$$

Manuscript received March 8, 1994; revised September 16, 1994. This work was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Air Force Office of Scientific Research.

The authors are with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA.

IEEE Log Number 9408794.

for time updates

$$P_{i+1}^{-1} \hat{x}_{i+1} = (I + F_i^{-*} P_{i|}^{-1} F_i^{-1} G_i Q_i G_i^*)^{-1} (F_i^{-*} P_{i|}^{-1} \hat{x}_{i|}),$$

$$P_{i+1}^{-1} = F_i^{-*} P_{i|}^{-1} F_i^{-1} \{I - G_i Q_i G_i^* F_i^{-*} P_{i|}^{-1} F_i^{-1}\}$$

where $\bar{Q}_i \triangleq Q_i - G_i Q_i P_{i+1}^{-1} G_i Q_i$. For convenience, we shall define $K_{p,i} \triangleq F_i P_i H_i^* R_i^{-1}$, $\bar{K}_{p,i} \triangleq K_{p,i} R_i^{1/2}$, $\bar{K}_{f,i} \triangleq K_{f,i} R_i^{1/2}$, and $\bar{K}_{b,i} \triangleq F_i^{-*} P_{i|}^{-1} F_i^{-1} G_i Q_i^{1/2}$.

Square-root algorithms for Kalman filtering can be nicely organized by combining what we call the covariance and information-form updates [22]. Let us first consider the basic so-called square-root covariance filter (SRCF) algorithm for measurement updates. Form the pre-array \mathcal{A} and block triangularize it by applying any unitary rotation Θ so that $\mathcal{A}\Theta$ has the form

$$\mathcal{A} = \begin{bmatrix} R_i^{1/2} & H_i P_i^{1/2} \\ 0 & P_i^{1/2} \\ -y_i^* R_i^{-*/2} & \hat{x}_i^* P_i^{-*/2} \end{bmatrix} \quad \mathcal{A}\Theta = \begin{bmatrix} X & 0 \\ Y & Z \end{bmatrix}$$

We may regard this as saying that the rotation Θ sets up a conformal (i.e., a norm- and angle-preserving) mapping between the (block) rows of \mathcal{A} and the rows of the post-array. An inner-product of the first row and a cross-product between the first and second rows yield the relations $R_i + H_i P_i H_i^* = X X^*$ and $H_i P_i = X Y^*$, from which we can identify $X = R_i^{1/2}$ and $Y = \bar{K}_{f,i}$. From the second block rows we obtain $P_i = \bar{K}_{f,i} \bar{K}_{f,i}^* + Z Z^*$, so that we can identify $Z = P_i^{1/2}$. Forming a cross-product between the first row and the third row, we obtain the relation that $y_i - H_i \hat{x}_i = X \alpha$, and verify that $\alpha = R_i^{-1/2} e_i$. Similarly, forming a cross-product between the second row and the third row gives us $\hat{x}_i = Y \alpha + Z \beta$. Therefore, we can select Z as $P_i^{-1/2} \hat{x}_{i|}$.

By inverting and transposing the first two blocks of the \mathcal{A} , we can form another basic so-called square-root information filter (SRIF) algorithm for measurement updates and combine the SRCF and the SRIF into a nice array form shown in the measurement-update equation of the following algorithm.

As far as time updates are concerned, we can form the following pre-array

$$\mathcal{A} = \begin{bmatrix} F_i P_{i|}^{1/2} & G_i Q_i^{1/2} \\ 0 & Q_i^{1/2} \\ \hat{x}_{i+1}^* P_{i+1}^{-*/2} & 0 \end{bmatrix}$$

and, after identifying the post-array by performing inner- or cross-products of the pre-array and comparing them, we can obtain the result shown in the time-update equation below.

Algorithm II.1—Square-Root Covariance and Information Filtering (Two-Array Form): Assume that $R_i > 0$, $\Pi_0 > 0$, and the F_i are invertible. Given $P_0^{-*/2} = \Pi_0^{-*/2}$, $P_0^{1/2} = \Pi_0^{1/2}$, and $P_0^{-*/2} \hat{x}_0 = 0$, we can write the following.

Measurement Updates:

$$\begin{bmatrix} R_i^{1/2} & H_i P_i^{1/2} \\ 0 & P_i^{1/2} \\ -H_i^* R_i^{-*/2} & P_i^{-*/2} \\ -y_i^* R_i^{-*/2} & \hat{x}_i^* P_i^{-*/2} \end{bmatrix} \Theta = \begin{bmatrix} R_{e,i}^{1/2} & 0 \\ \bar{K}_{f,i} & P_{i|}^{1/2} \\ R_{e,i}^{-*/2} & -R_i^{-1} H_i P_{i|}^{1/2} \\ 0 & P_{i|}^{-*/2} \\ -e_i^* R_{e,i}^{-*/2} & \hat{x}_{i|}^* P_{i|}^{-*/2} \end{bmatrix}$$

where Θ is any unitary rotation that zeros out the (1, 2) or (4, 1) entry of the pre-array.

Time Updates:

$$\begin{bmatrix} F_i P_{i|}^{1/2} & G_i Q_i^{1/2} \\ 0 & Q_i^{1/2} \\ -F_i^{-*} P_{i|}^{-*/2} & 0 \\ -G_i^* F_i^{-*} P_{i|}^{-*/2} & Q_i^{-*/2} \\ \hat{x}_{i+1}^* P_{i+1}^{-*/2} & 0 \end{bmatrix} \Theta = \begin{bmatrix} P_{i+1}^{1/2} & 0 \\ Q_i G_i P_{i+1}^{-*/2} & \bar{Q}_i^{1/2} \\ P_{i+1}^{-*/2} & -\bar{K}_{b,i} \\ 0 & \bar{Q}_i^{-*/2} \\ \hat{x}_{i+1}^* P_{i+1}^{-*/2} & -\hat{x}_{i+1}^* \bar{K}_{b,i} \end{bmatrix}$$

where Θ is any unitary rotation that zeros out the (1, 2) or (4, 1) entry of the pre-array.

Estimate Construction: The filtered estimates \hat{x}_{i+1} can be constructed from the post-array of the time-update equation in Algorithm II.1 as $\hat{x}_{i+1} = (P_{i+1}^{1/2})^{-1} (P_{i+1}^{-1/2} \hat{x}_{i+1})$ or from the post-array of the measurement-update equation in Algorithm II.1 as $\hat{x}_{i+1} = (F_i) \{(\hat{x}_i) + (\bar{K}_{f,i}) (R_{e,i}^{-1/2} e_i)\}$.

The following corollary shows the square-root covariance and information filtering (SRCIF) algorithm in a single-array form, which is useful when the input and output noises are correlated; see [22]. We shall use this single-array form in Section III to implement the Bryson-Frazier smoothing formula in square-root form.

Corollary II.1—Square-Root Covariance and Information Filtering (Single-Array Form): Assume that $R_i > 0$, $\Pi_0 > 0$, and the F_i are invertible. Given $P_0^{-*/2} = \Pi_0^{-*/2}$, $P_0^{1/2} = \Pi_0^{1/2}$, and $P_0^{-*/2} \hat{x}_0 = 0$

$$\begin{bmatrix} R_i^{1/2} & H_i P_i^{1/2} & 0 \\ 0 & F_i P_i^{1/2} & G_i Q_i^{1/2} \\ -F_i^{-*} H_i^* R_i^{-*/2} & F_i^{-*} P_i^{-*/2} & 0 \\ G_i^* F_i^{-*} H_i^* R_i^{-*/2} & -G_i^* F_i^{-*} P_i^{-*/2} & Q_i^{-*/2} \\ -y_i^* R_i^{-*/2} & \hat{x}_i^* P_i^{-*/2} & 0 \end{bmatrix} = \begin{bmatrix} R_{e,i}^{1/2} & 0 & 0 \\ \bar{K}_{p,i} & P_{i+1}^{1/2} & 0 \\ 0 & P_{i+1}^{-*/2} & -\bar{K}_{b,i} \\ 0 & 0 & \bar{Q}_i^{-*/2} \\ -e_i^* R_{e,i}^{-*/2} & \hat{x}_{i+1}^* P_{i+1}^{-*/2} & -\hat{x}_{i+1}^* \bar{K}_{b,i} \end{bmatrix}$$

where Θ is any unitary rotation that either lower-triangularizes the first two (block) rows or upper-triangularizes the third and fourth (block) rows of the pre-array. The filtered estimates \hat{x}_{i+1} can be constructed via $\hat{x}_{i+1} = (P_{i+1}^{1/2})^{-1} (P_{i+1}^{-1/2} \hat{x}_{i+1})$ or $\hat{x}_{i+1} = F_i(\hat{x}_i) + \bar{K}_{p,i} (R_{e,i}^{-1/2} e_i)$.

III. BRYSON-FRAZIER FORMULAS AND THEIR SQUARE-ROOT ALGORITHMS

The fixed-interval smoothing problem is to find $\{\hat{x}_{i|N}\}_{i=0:N}$ given output data $\{y_j\}_{j=0:N}$. In 1963, Bryson and Frazier [21] suggested two smoothing formulas that use the $\{\hat{x}_i\}$ or the $\{\hat{x}_{i|}\}$ of Kalman filtering, and certain quantities defined by the recursions

$$\lambda_{i|N} = F_{p,i}^* \lambda_{i+1|N} + H_i^* R_{e,i}^{-1} \lambda_{N+1|N} = 0, \quad (1)$$

$$\Lambda_{i|N} = F_{p,i}^* \Lambda_{i+1|N} F_{p,i} + H_i^* R_{e,i}^{-1} H_i \Lambda_{N+1|N} = 0 \quad (2)$$

where $F_{p,i} \triangleq F_i - K_{p,i} H_i$. (See the Appendix for the brief derivation.)

Algorithm III.1—BF Smoothing Formulas:

Case 1) (based on \hat{x}_i) $\hat{x}_{i|N} = \hat{x}_i + P_i \lambda_{i|N}$,

$$P_{i|N} = P_i - P_i \Lambda_{i|N} P_i.$$

Case 2) (based on $\hat{x}_{i|}$) $\hat{x}_{i|N} = \hat{x}_{i|} + P_{i|} F_i^* \lambda_{i+1|N}$,

$$P_{i|N} = P_{i|} - P_{i|} F_i^* \Lambda_{i+1|N} F_i P_{i|}.$$

Square-Root BF Smoothing Algorithms

The coefficients $F_{p,i}^*$ and $H_i^* R_{i,i}^{-1/2}$ of (1) can be computed simultaneously when the P_i are computed via the conventional equation-based algorithm. Similarly, we can get some terms related to the coefficients when we use a square-root algorithm for propagating the P_i . Consider the following array, formed from the array of Corollary II.1 extended with one more row

$$\begin{bmatrix} R_i^{1/2} & H_i P_i^{1/2} & 0 \\ 0 & F_i P_i^{1/2} & G_i Q_i^{1/2} \\ 0 & I & 0 \\ -y_i^* R_{i,i}^{-*/2} & \hat{x}_i^* P_i^{-*/2} & 0 \end{bmatrix} \Theta = \begin{bmatrix} R_{i,i}^{1/2} & 0 & 0 \\ \bar{K}_{p,i} & P_{i+1}^{1/2} & 0 \\ X_i & Y_i & (*) \\ -c_i^* R_{i,i}^{-*/2} & \hat{x}_{i+1}^* P_{i+1}^{-*/2} & (*) \end{bmatrix}$$

Cross-products between the first and the third rows and between the second and the third ones yield the equation $X_i = P_i^{-1/2} (P_i H_i^* R_{i,i}^{-*/2})$, $Y_i = P_i^{*/2} (F_i - \bar{K}_{p,i} H_i)^* P_{i+1}^{-*/2} = P_i^{*/2} F_{p,i}^* P_{i+1}^{-*/2}$. Rewriting (1) using the $\{X_i\}$ and $\{Y_i\}$ gives the formulas

$$(P_i^{*/2} \lambda_{i|N}) = (Y_i) (P_{i+1}^{*/2} \lambda_{i+1|N}) + (X_i) (R_{i,i}^{-1/2} e_i)$$

which recursively propagates $\{P_i^{*/2} \lambda_{i|N}\}$ instead of $\{\lambda_{i|N}\}$. With the $\{P_i^{-1/2} \hat{x}_i\}$ recursively found from the above array, we can use the Bryson-Frazier formulas to obtain

$$\hat{x}_{i|N} = (P_i^{1/2}) \{ (P_i^{-1/2} \hat{x}_i) + (P_i^{*/2} \lambda_{i|N}) \}. \quad (3)$$

It is impossible, however, to find the quantities $\{\lambda_{i|N}\}$ in (2) by directly using the $\{X_i\}$ and $\{Y_i\}$. Therefore, we shall separate the $\{P_i^{*/2} \lambda_{i|N}\}$ into $\{P_i^{*/2} \Lambda_{i|N}^{1/2}\}$ and $\{\Lambda_{i|N}^{-1/2} \lambda_{i|N}\}$ and form another array using the $\{X_i\}$, $\{Y_i\}$ and $\{P_i^{*/2} \Lambda_{i|N}^{1/2}\}$

$$\begin{bmatrix} X_i & (Y_i) (P_{i+1}^{*/2} \Lambda_{i+1|N}^{1/2}) \\ c_i^* R_{i,i}^{-*/2} & (\lambda_{i+1|N}^* \Lambda_{i+1|N}^{-*/2}) \end{bmatrix} \bar{\Theta} = \begin{bmatrix} W \\ \mu \end{bmatrix}$$

where $\bar{\Theta}$ is any unitary operator that zeros out the (1, 2) entry of the pre-array. Inner- or cross-products of the array entries yield

$$W W^* = X_i X_i^* + (Y_i) (P_{i+1}^{*/2} \Lambda_{i+1|N}^{1/2}) (\Lambda_{i+1|N}^{*/2} P_{i+1}^{1/2}) (Y_i)^*, \\ \mu = W^{-1} \{ (X_i) (R_{i,i}^{-1/2} e_i) + (Y_i) (\Lambda_{i+1|N}^{-1/2} \lambda_{i+1|N}) \}.$$

Hence, the new array recursively provides $\{W = P_i^{*/2} \Lambda_{i|N}^{1/2}\}$ and $\{\mu = \Lambda_{i|N}^{-1/2} \lambda_{i|N}\}$, so that the $\{P_{i|N}\}$ can then be found as

$$P_{i|N} = (P_i^{1/2}) \{ I - (P_i^{*/2} \Lambda_{i|N}^{1/2}) (\Lambda_{i|N}^{*/2} P_i^{1/2}) \} (P_i^{*/2}). \quad (4)$$

These results are summarized in the following proposition.

Proposition III.1—Square-Root BF Smoothing (Case 1): Assume that $R_i > 0$. Perform Step 1 first, and then Step 2 and Step 3, together.

Step 1: For $(P_0^{-1/2} \hat{x}_0) = 0$ and $(P_0^{1/2}) = \Pi_0^{1/2}$, propagate $\{P_i^{-1/2} \hat{x}_i\}$ and $\{P_i^{1/2}\}$ using a forward recursion

$$\begin{bmatrix} R_i^{1/2} & H_i P_i^{1/2} & 0 \\ 0 & F_i P_i^{1/2} & G_i Q_i^{1/2} \\ 0 & I & 0 \\ -y_i^* R_{i,i}^{-*/2} & \hat{x}_i^* P_i^{-*/2} & 0 \end{bmatrix} \Theta = \begin{bmatrix} R_{i,i}^{1/2} & 0 & 0 \\ \bar{K}_{p,i} & P_{i+1}^{1/2} & 0 \\ X_i & Y_i & (*) \\ -c_i^* R_{i,i}^{-*/2} & \hat{x}_{i+1}^* P_{i+1}^{-*/2} & (*) \end{bmatrix}$$

where Θ is any unitary operator that lower-triangularizes the first and second rows of the pre-array and $(*)$ indicates the redundant entries. During this forward recursion, generate and save the following variables: for Step 2, $\{P_i^{*/2} H_i^* R_{i,i}^{-*/2}\}$, $\{P_i^{*/2} F_{p,i}^* P_{i+1}^{-*/2}\}$, $\{R_{i,i}^{-1/2} e_i\}$, and for Step 3, $\{P_i^{1/2}\}$, $\{P_i^{-1/2} \hat{x}_i\}$.

Step 2: For $(\Lambda_{N+1|N}^{-1/2} \lambda_{N+1|N}) = 0$ and $(P_{N+1|N}^{*/2} \Lambda_{N+1|N}^{1/2}) = 0$, propagate $\{\Lambda_{i|N}^{-1/2} \lambda_{i|N}\}$ and $\{P_i^{*/2} \Lambda_{i|N}^{1/2}\}$ using a backward recursion

$$\begin{bmatrix} (P_i^{*/2} H_i^* R_{i,i}^{-*/2}) & (P_i^{*/2} F_{p,i}^* P_{i+1}^{-*/2}) (P_{i+1}^{*/2} \Lambda_{i+1|N}^{1/2}) \\ (c_i^* R_{i,i}^{-*/2}) & (\lambda_{i+1|N}^* \Lambda_{i+1|N}^{-*/2}) \end{bmatrix} \bar{\Theta} = \begin{bmatrix} (P_i^{*/2} \Lambda_{i|N}^{1/2}) & 0 \\ (\lambda_{i|N}^* \Lambda_{i|N}^{-*/2}) & (*) \end{bmatrix}$$

where $\bar{\Theta}$ is any unitary operator that zeros out the (1, 2) entry of the pre-array.

Step 3: Calculate smoothed estimates and their error covariances in parallel with Step 2 via (3) and (4).

If we are interested only in smoothed estimates but not in their error covariances, then we can speed up the algorithm and reduce the amount of storage.

Proposition III.2—Square-Root BF Smoothing (Estimates Only):

Step 1: For $(P_0^{-1/2} \hat{x}_0) = 0$ and $(P_0^{1/2}) = \Pi_0^{1/2}$, propagate $\{P_i^{-1/2} \hat{x}_i\}$ and $\{P_i^{1/2}\}$ using the same array as in Step 1 in Proposition III.1. Save the following: $\{P_i^{1/2}\}$, $\{P_i^{-1/2} \hat{x}_i\}$, $\{P_i^{*/2} F_{p,i}^* P_{i+1}^{-*/2}\}$, $\{c_i = (P_i^{*/2} H_i^* R_{i,i}^{-*/2}) (R_{i,i}^{-1/2} e_i)\}$.

Step 2: For $(P_{N+1|N}^{*/2} \lambda_{N+1|N}) = 0$, propagate $\{P_i^{*/2} \lambda_{i|N}\}$ using the following backward equation

$$(P_i^{*/2} \lambda_{i|N}) = (P_i^{*/2} F_{p,i}^* P_{i+1}^{-*/2}) (P_{i+1}^{*/2} \lambda_{i+1|N}) + c_i.$$

Step 3: Calculate smoothed estimates as

$$\hat{x}_{i|N} = (P_i^{1/2}) \{ (P_i^{-1/2} \hat{x}_i) + (P_i^{*/2} \lambda_{i|N}) \}.$$

Remark 1: The main advantages of the square-root BF algorithm are that there is neither matrix inversion nor back substitution, and all gains for forward and backward recursions can be computed via a forward pass only. Therefore, whereas the forward pass procedure requires a large computation, the backward pass procedure requires relatively small computation. As disadvantages, the error covariance $P_{i|N}$ might not be positive semi-definite due to round-off errors and, during the procedures, we should save $\{P_i^{*/2} H_i^* R_{i,i}^{-*/2}\}$, $\{P_i^{*/2} F_{p,i}^* P_{i+1}^{-*/2}\}$, $\{R_{i,i}^{-1/2} e_i\}$, $\{P_i^{1/2}\}$, $\{P_i^{-1/2} \hat{x}_i\}$, which requires memory of the order of $(N+1)np$, $(N+1)n^2$, $(N+1)p$, $(N+1)n(n+1)/2$, $(N+1)n$, respectively. The major costs come from $\{P_i^{*/2} F_{p,i}^* P_{i+1}^{-*/2}\}$ and $\{P_i^{1/2}\}$.

If we are interested in $\hat{x}_{i|N}$ but not in $P_{i|N}$ and the F_i are nonsingular, we can find another square-root form that compensates for the latter drawback. Let us think about a method for propagating $\{\hat{x}_i\}$ and $\{P_i \lambda_{i|N}\}$. The first quantities can be composed from the array in Step 1 of Proposition III.1 via $(P_i^{1/2}) (P_i^{-1/2} \hat{x}_i)$. For the second quantities, let us consider the following variation of (1): $(P_i \lambda_{i|N}) = F_{s,i} (P_{i+1} \lambda_{i+1|N}) + (\bar{K}_{f,i}) (R_{i,i}^{-1/2} e_i)$, where $F_{s,i}$ denotes

$$P_i F_{p,i}^* P_{i+1}^{-1} = F_i^{-1} (I - G_i Q_i G_i^* P_{i+1}^{-1}) = F_i^{-1} (I - G_i \bar{Q}_i^{1/2} \bar{K}_{b,i}^*),$$

Since the two equations in Algorithm II.1 give $\{\bar{K}_{f,i}\}$, $\{R_{i,i}^{-1/2} e_i\}$, $\{\bar{K}_{b,i}\}$, and $\{\bar{Q}_i^{1/2}\}$, we can find $\{P_i \lambda_{i|N}\}$ via

$$(P_i \lambda_{i|N}) = (F_i^{-1}) \{ I - G_i (\bar{Q}_i^{1/2}) (\bar{K}_{b,i}^*) \} (P_{i+1} \lambda_{i+1|N}) + (\bar{K}_{f,i}) (R_{i,i}^{-1/2} e_i). \quad (5)$$

Proposition III.3—Square-Root BF Smoothing (Case 2): Assume that $R_i > 0$ and the F_i are nonsingular.

Step 1: For $(P_0^{-1/2} \hat{x}_0) = 0$ and $(P_0^{1/2}) = \Pi_0^{1/2}$, propagate $\{(P_i^{-1/2} \hat{x}_i)\}$ and $\{(P_i^{1/2})\}$ using measurement- and time-update equations.

Measurement Updates:

$$\begin{bmatrix} 0 & P_i^{1/2} \\ -H_i^* R_i^{-*/2} & P_i^{-*/2} \\ -y_i^* R_i^{-*/2} & \hat{x}_i^* P_i^{-*/2} \end{bmatrix} \Theta = \begin{bmatrix} \bar{K}_{f,i} & P_{f,i}^{1/2} \\ 0 & P_{f,i}^{-*/2} \\ -\bar{c}_i^* R_{f,i}^{-*/2} & \hat{x}_{f,i}^* P_{f,i}^{-*/2} \end{bmatrix}$$

where Θ is any unitary operator that zeros out the (2, 1) entry of the pre-array. Save $\{(\bar{K}_{f,i})(R_{f,i}^{-1/2} \bar{c}_i)\}$.

Time Updates:

$$\begin{bmatrix} F_i P_{f,i}^{1/2} & G_i Q_{f,i}^{1/2} \\ 0 & Q_{f,i}^{1/2} \\ F_i^{-*/2} P_{f,i}^{-*/2} & 0 \\ \hat{x}_{f,i}^* P_{f,i}^{-*/2} & 0 \end{bmatrix} \Theta = \begin{bmatrix} P_{i+1}^{1/2} & 0 \\ Q_{i+1}^* P_{i+1}^{-1/2} & \bar{Q}_{i+1}^{1/2} \\ P_{i+1}^{-*/2} & -\bar{K}_{b,i} \\ \hat{x}_{i+1}^* P_{i+1}^{-*/2} & -\hat{x}_{i+1}^* \bar{K}_{b,i} \end{bmatrix}$$

where Θ is any unitary operator that zeros out the (1, 2) entry of the pre-array. Save the following: $\{(P_i^{1/2})(P_i^{-1/2} \hat{x}_i)\}$, $\{(\bar{Q}_{i+1}^{1/2})(\bar{K}_{b,i}^*)\}$.

Step 2: For $(P_{N+1|N} \lambda_{N+1|N}) = 0$, propagate $(P_i \lambda_{i|N})$ via (5).

Step 3: Calculate smoothed estimates via $\hat{x}_{i|N} = \{(P_i^{-1/2} \hat{x}_i)\} + (P_i \lambda_{i|N})$. ■

Remark 2: An interesting feature of this algorithm is that the SRIF block in the measurement-update equation in Algorithm II.1 is chosen for measurement updates, whereas the SRCF block in the time-update equation in Algorithm II.1 is chosen for time updates.

Remark 3: When the F_i are time-variant, there is no difference of memory size between the two propositions if we save the P_i by writing over the F_i . Otherwise, the memory size is quite different. The latter case requires memory of the order of $(N+1)n$, $(N+1)n$, $(N+1)n$ corresponding to $\{(P_i^{1/2})(P_i^{-1/2} \hat{x}_i)\}$, $\{(\bar{K}_{f,i})(R_{f,i}^{-1/2} \bar{c}_i)\}$ and $\{(\bar{Q}_{i+1}^{1/2})(\bar{K}_{b,i}^*)\}$, respectively.

Fast Square-Root BF Smoothing Algorithms

Now, we shall derive a fast square-root (or so-called Chandrasekhar-type) form for the BF formulas. If the system matrices are constant, we can apply a fast square-root algorithm for forwards Kalman filtering, which will produce the normalized Kalman gain \bar{K}_p , and the square-root innovation covariance $R_{f,i}^{1/2}$. Recall that we generated some quantities from a forward pass using the array in Proposition III.3: $\{(\bar{K}_{f,i})(R_{f,i}^{-1/2} \bar{c}_i)\}$ and $\{(\bar{Q}_{i+1}^{1/2})(\bar{K}_{b,i}^*)\}$, which are needed to form $F_{*,i} \triangleq F^{-1}(I - G\bar{Q}_{i+1}^{1/2}\bar{K}_{b,i}^*)$ for the backward pass. Unfortunately, the term $F_{*,i}$ cannot be found using the fast square-root covariance algorithm for forwards Kalman filtering, because this term contains P_{i+1}^{-1} which is impossible to obtain through the displacement of the P_i . Therefore, we must employ one more fast square-root algorithm related to the displacement of the P_{i+1}^{-1} .

Fast Square-Root Covariance Filtering. We first recall that the fast square-root form uses matrices $\{L_i\}$ or $\{N_i\}$ defined as $L_i J_i L_i^* \triangleq P_{i+1} - P_i$ or $N_i J_i N_i^* \triangleq P_{i+1|+1} - P_{i|+}$. Consider the following array

$$\begin{bmatrix} R_{f,i}^{1/2} & H L_i \\ \bar{K}_{f,i} & F L_i \end{bmatrix} \Theta = \begin{bmatrix} X & 0 \\ Y & Z \end{bmatrix}, J = \begin{bmatrix} I & 0 \\ 0 & J_L \end{bmatrix}$$

where Θ is any J -unitary operator that zeros out the (1, 2) entry of the pre-array. J -inner or J -cross products of the array entries yield $XX^* = R_{f,i+1}$, $YX^* = \bar{K}_{f,i+1} R_{f,i+1}^{1/2}$, $ZJ_L Z^* = P_{i+2} - P_{i+1} =$

$L_{i+1} J_L L_{i+1}^*$. Therefore, we can identify $X = R_{f,i+1}^{1/2}$, $Y = \bar{K}_{f,i+1}$, $Z = L_{i+1}$.

If we want to find the $\bar{K}_{f,i}$ from this array, we should use N_i rather than L_i , which obey the interesting formula $L_{i+1} \Theta = F N_i$, where Θ is some J_L -unitary operator. For a proof, note that, since $P_{i+1} = F P_{i|+} F^* + G Q G^*$ and thus $L_{i+1} J_L L_{i+1}^* = P_{i+2} - P_{i+1} = F(P_{i+1|+1} - P_{i|+})F^* = F N_i J_i N_i^* F^*$, therefore, the above array form can be rewritten by

$$\mathcal{F} \begin{bmatrix} R_{f,i}^{1/2} & H F N_{i-1} \\ \bar{K}_{f,i} & F N_{i-1} \end{bmatrix} \Theta = \mathcal{F} \begin{bmatrix} R_{f,i+1}^{1/2} & 0 \\ \bar{K}_{f,i+1} & N_i \end{bmatrix}$$

where $\mathcal{F} = I - F$ and Θ is any J -unitary operator that zeros out the (1, 2) entry of the pre-array. Hence, unless the matrix F is singular, we can generate the $\{(\bar{K}_{f,i})(R_{f,i}^{1/2})^{-1}\}$ from the fast square-root covariance algorithm. Now, to construct the $F_{*,i}$, we are going to provide a fast square-root information algorithm.

Fast Square-Root Information Filtering. The objective of the algorithm is to find $\Gamma_{*,i} = F^{-1}(I - G\bar{Q}_{i+1}^{1/2}\bar{K}_{b,i}^*)$, and indeed, $\{(\bar{Q}_{i+1}^{1/2})(\bar{K}_{b,i}^*)\}$. Consider the following array

$$\begin{bmatrix} \bar{Q}_{i+1}^{*/2} & G^* F^{-*} M_i \\ \bar{K}_{b,i} & F^{-*} M_i \end{bmatrix} \Theta = \begin{bmatrix} X & 0 \\ Y & Z \end{bmatrix}, J = \begin{bmatrix} I & 0 \\ 0 & J_M \end{bmatrix}$$

where Θ is any J -unitary operator that zeros out the (1, 2) entry of the pre-array, and $M_i J_M M_i^* \triangleq P_{i+1}^{-1} - P_i^{-1}$. Forming J -inner or J -cross products of the array entries yields

$$XX^* = Q + G^* F^{-*} (P_{i+1}^{-1} + H^* R^{-1} H) F^{-1} G = Q_{i+1}^{-1},$$

$$YX^* = F^{-*} (P_{i+1}^{-1} + H^* R^{-1} H) F^{-1} G = K_{b,i+1} Q_{i+1}^{1/2},$$

$$ZJ_M Z^* = P_{i+2}^{-1} - P_{i+1}^{-1} = M_{i+1} J_M M_{i+1}^*.$$

Therefore, we can identify $X = Q_{i+1}^{*/2}$, $Y = K_{b,i+1}$, $Z = M_{i+1}$. Here we note that the minimal rank of the J_L is as same as that of the J_M . To show this, consider initial conditions M_0 and N_0 . Then note that since $P_1 - P_0 \triangleq L_0 J_L L_0^* = F^{-1} N_{-1} J_L N_{-1}^* F^{-1}$, $P_1^{-1} - P_0^{-1} = -P_0^{-1} L_0 (J_L + L_0^* P_0^{-1} L_0)^{-1} L_0^* P_0^{-1} (\triangleq M_0 J_M M_0^*)$. Based on these results, we can present a fast square-root smoothing algorithm.

Proposition III.4

Step 1 (Fast Square-Root Covariance Filtering)

$$\begin{bmatrix} R_{f,i}^{1/2} & H F N_{i-1} \\ \bar{K}_{f,i} & F N_{i-1} \\ R_{f,i}^{-*/2} & 0 \end{bmatrix} \Theta = \begin{bmatrix} R_{f,i+1}^{1/2} & 0 \\ \bar{K}_{f,i+1} & N_i \\ R_{f,i+1}^{-*/2} & (*) \end{bmatrix}$$

where $J = I - J_L$ and Θ is any J -unitary operator which zeros out the (1, 2) entry of the pre-array and

$$L_0 J_L L_0^* = P_1 - P_0, L_0 = F N_{-1},$$

$$R_{f,0}^{1/2} = (R + H P_0 H^*)^{1/2}, \bar{K}_{f,0} = P_0 H^* R_{f,0}^{-*/2}.$$

For $\hat{x}_0 = 0$, compute $\{\mu_i\}$ and $\{x_i\}$ via $\mu_i = \bar{K}_{f,i} R_{f,i}^{*/2} (y_i - H \hat{x}_i)$ and $\hat{x}_{i+1} = F(\hat{x}_i + \mu_i)$, respectively, and save them.

Step 2 (Fast Square-Root Information Filtering)

$$\begin{bmatrix} \bar{Q}_{i+1}^{*/2} & G^* F^{-*} M_i \\ \bar{K}_{b,i} & F^{-*} M_i \\ \bar{Q}_{i+1}^{1/2} & 0 \end{bmatrix} \Theta = \begin{bmatrix} \bar{Q}_{i+1}^{*/2} & 0 \\ \bar{K}_{b,i+1} & M_{i+1} \\ \bar{Q}_{i+1}^{1/2} & (*) \end{bmatrix}$$

where $J = I - J_M$ and Θ is any J -unitary operator that zeros out the (1, 2) entry of the pre-array and

$$M_0 J_M M_0^* = -P_0^{-1} L_0 (J_L + L_0^* P_0^{-1} L_0)^{-1} L_0^* P_0^{-1},$$

$$\bar{Q}_0^{*/2} = (Q - Q G P_1^{-1} G^* Q)^{-*/2}, \bar{K}_{b,0} = P_1^{-1} G Q \bar{Q}_0^{*/2}.$$

During the procedure, save $\{(\bar{Q}_{i+1}^{1/2})(\bar{K}_{b,i}^*)\}$. For $(P_{N+1} \lambda_{N+1|N}) = 0$, propagate

$$(P_i \lambda_{i|N}) = F^{-1} \{I - G\{(\bar{Q}_{i+1}^{1/2})(\bar{K}_{b,i}^*)\}\} (P_{i+1} \lambda_{i+1|N}) + \mu_i.$$

TABLE I
CONSTRAINTS

SR algorithms	Error Cov.	F_i	System
Prop. III.1	yes		
Prop. III.2	no		
Prop. III.3	no	invertible	
Prop. III.4	no	invertible	constant

TABLE II
COMPARISONS

SR algorithms	Storage	Speed	Stability
Prop. III.1	$O(\frac{3}{2}(N+1)n^2)$	slowest	Good
Prop. III.2	$O(\frac{3}{2}(N+1)n^2)$	slow	Good
Prop. III.3	$O((N+1)np)$	slow	Good
Prop. III.4	$O((N+1)np)$	fastest	(?)

Step 3 (Smoothed Estimates): Compute the smoothed estimates as $\hat{x}_{i|N} = \hat{x}_i + (P_i \lambda_{i|N})$.

Remark 4: This algorithm requires memory of order of $(N+1)n$, $(N+1)n$, $(N+1)np$ for the $\{\hat{x}_i\}$, $\{\mu_i\}$, and $\{(\hat{Q}_i^{1/2})(\hat{K}_{b,i}^*)\}$.

Remark 5 (Comparisons): If error covariances need to be computed, only Proposition III.1 is available. If the matrices F_i are not nonsingular, Propositions III.1 and III.2 are possible, where using the latter is faster than using the former. If the F_i are nonsingular and error covariances do not need to be computed, Propositions III.1, III.2, and III.3 can be used, where Proposition III.3 gives better performance than the other propositions because of less storage and computation. If, in addition, the system matrices are constant, all the propositions can be used. In this case, we recommend Proposition III.3 and Proposition III.4, depending on a trade-off between speed and stability—see Tables I and II.

IV. CONCLUDING REMARKS

The new square-root algorithms use arrays including filtered or smoothed estimates and their covariances and avoid inversion or back substitution, which are major advantages over previously known algorithms with respect to systolic array and parallel implementations as well as numerical stability and conditioning. Moreover, the square-root array form supplies flexibility in implementing the rotation in various ways.

For constant-parameter systems with invertible F , a fast square-root (or Chandrasekhar-type) smoothing algorithm is suggested, which has advantages of small storage and fast computation over the others. But the numerical properties of this algorithm need further investigation because it uses J -unitary rotations rather than unitary rotations.

APPENDIX

BRIEF DERIVATION OF BF FORMULAS

The fixed-interval smoothing problem is to find $\{\hat{x}_{i|N}\}_{i=0:N}$ given output data $\{y_j\}_{j=0:N}$. In this case, since the smoothed estimate can be rewritten with the innovations e_j so that $\hat{x}_{i|N} = \sum_{j=0}^N E(x_i e_j^*) E^{-1}(e_j e_j^*) e_j$, where " $E(x)$ " denotes the expectation of x , we have the following two expressions of $\hat{x}_{i|N}$

$$\sum_{j=0}^{i-1} E(x_i e_j^*) E^{-1}(e_j e_j^*) e_j + \sum_{j=i}^N E(x_i e_j^*) E^{-1}(e_j e_j^*) e_j$$

or

$$\sum_{j=0}^i E(x_i e_j^*) E^{-1}(e_j e_j^*) e_j + \sum_{j=i+1}^N E(x_i e_j^*) E^{-1}(e_j e_j^*) e_j$$

where $\sum_{j=0}^{i-1} E(x_i e_j^*) E^{-1}(e_j e_j^*) e_j$ can be denoted as \hat{x}_i and $\sum_{j=i}^N E(x_i e_j^*) E^{-1}(e_j e_j^*) e_j$, as $\hat{x}_{i|N}$. After performing complicated calculation, we can verify that, for $i \leq j$

$$E(x_i e_j^*) = P_i \Phi_p^*(j, i) H_j^* = P_{i|j} F_i^* \Phi_p^*(j, i+1) H_j^*$$

where $\Phi_p(j, i) = F_{p,j-1} F_{p,j-2} \cdots F_{p,i}$, $F_{p,i} = F_i - K_{p,i} H_i^*$, $K_{p,i} = F_i K_{f,i}$. Therefore, defining $\lambda_{i|N}$ as $\sum_{j=i}^N \Phi_p^*(j, i) H_j^* R_{e,j}^{-1} e_j$ gives the two recursive equations for the smoothed estimates in Algorithm III.1.

Since e_i is orthogonal to e_j for $i \neq j$, the two recursive equations for the smoothed estimates in Algorithm III.1 provide the relation

$$\begin{aligned} E(\hat{x}_{i|N} \hat{x}_i^*) &= E(\hat{x}_{i|N} \hat{x}_i^*) + P_{i|j} F_i^* E(\lambda_{i+1|N} \lambda_{i+1|N}^*) F_i P_{i|j} \\ &= E(\hat{x}_i \hat{x}_i^*) + P_i E(\lambda_{i|N} \lambda_{i|N}^*) P_i. \end{aligned}$$

Using the fact that $P_{i|N} = E(x_i x_i^*) - E(\hat{x}_{i|N} \hat{x}_{i|N}^*)$, $P_{i|j} = E(x_i x_i^*) - E(\hat{x}_{i|j} \hat{x}_{i|j}^*)$, and $P_i = E(x_i x_i^*) - E(\hat{x}_i \hat{x}_i^*)$, we can obtain the two recursive equations for the covariance matrices in Algorithm III.1.

REFERENCES

- [1] J. E. Potter and R. G. Stern, "Statistical filtering of space navigation measurements," in *Proc. 1963 AIAA Guidance Contr. Conf.*, 1963.
- [2] S. F. Schmidt, "Computational techniques in Kalman filtering," in *Theory and Applications of Kalman Filtering*, NATO Advisory Group for Aerospace Research and Development, Feb. 1970.
- [3] G. H. Golub, "Numerical methods for solving linear least squares problems," *Numerical Mathematics*, vol. 7, pp. 206–216, 1965.
- [4] P. Dyer and S. McReynolds, "Extension of square-root filtering to include process noise," *J. Optimization Theory Applications*, vol. 3, no. 6, pp. 444–459, 1969.
- [5] P. G. Kaminski, A. E. Bryson, and S. F. Schmidt, "Discrete square-root filtering—a survey of current techniques," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 727–736, 1971.
- [6] M. Morf and T. Kailath, "Square root algorithms for least squares estimation," *IEEE Trans. Automat. Contr.*, vol. AC-20, pp. 487–497, Aug. 1975.
- [7] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.
- [8] T. Kailath, *Lectures on Wiener and Kalman Filtering*, 2nd ed. New York: Springer-Verlag, 1981.
- [9] P. G. Kaminski, "Square-root filtering and smoothing for discrete processes," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1971.
- [10] G. J. Bierman, "A reformulation of the Rauch–Tung–Striebel discrete time fixed interval smoother," in *Proc. IEEE Conf. Decis. Contr.*, Austin, TX, Dec. 1988, pp. 840–844.
- [11] G. J. Bierman, "Sequential square-root filtering and smoothing discrete linear systems," *Automatica*, vol. 10, pp. 147–158, 1974.
- [12] K. Watanabe and D. S. G. Tzafestas, "New computationally efficient formula for backward-pass fixed-interval smoother and its UD factorization algorithm," *IEE Proc. D*, vol. 136, no. 2, pp. 73–78, 1989.
- [13] S. R. McReynolds, "Covariance factorization algorithms for fixed-interval smoothing of linear discrete dynamic systems," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 1181–1183, Oct. 1990.
- [14] K. Watanabe, "A new forward-pass fixed-interval smoother using the U-D information matrix factorization," *Automatica*, vol. 22, no. 4, pp. 465–475, 1986.
- [15] J. R. Dobbins, "Covariance factorization techniques for least squares estimation," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1979.
- [16] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1980.
- [17] S. Chandrasekhar, "On the radiative equilibrium of a stellar atmosphere, Pt. XXI," *Astrophys. J.*, vol. 106, pp. 152–216, 1947, Pt. XXII, vol. 107, pp. 48–72, 1948.
- [18] S. Chandrasekhar, *Radiative Transfer*. New York: Dover Publications, 1960.
- [19] T. Kailath, "Some new algorithms for recursive estimation in constant linear systems," *IEEE Trans. Inform. Theory*, vol. 19, pp. 750–760, Nov. 1973.
- [20] M. Morf, G. S. Sidhu, and T. Kailath, "Some new algorithms for recursive estimation in constant, linear, discrete-time systems," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 315–323, Aug. 1974.

- [21] A. E. Bryson and M. Frazier, "Smoothing for linear and nonlinear dynamic systems," *Aero. Syst. Div.*, pp. 353–364, 1963. Wright-Patterson Air Force Base, OH, TDR 63-119.
- [22] P. Park and T. Kailath, "New square-root algorithms for Kalman filtering," *IEEE Trans. Automat. Contr.*, vol. 40, July 1995.

Robust H_∞ Control of Uncertain Nonlinear System via State Feedback

Tielong Shen and Katsutoshi Tamura

Abstract—This technical note presents a solution of robust H_∞ control problem for an affine nonlinear system with gain bounded uncertainty. It is shown that the L_2 -induced norm of a closed-loop system with the uncertainty is less than one if an extended nonlinear system with no uncertainty is dissipative with respect to a supply rate. In consequence of this result, a state feedback law such that the closed-loop system has H_∞ robust performance is derived based on the Hamilton–Jacobi inequality. It is also shown that the existing results for linear systems are special cases of the presented results.

I. INTRODUCTION

A popular method to solve the standard H_∞ control problem is the state-space approach developed in [1] which is based on the relation between the H_∞ norm of linear system and the Riccati equation or inequality. Recently, attention was focused to robust H_∞ performance problem and the state-space approach has been extended to solve this problem for linear systems with parameter perturbation [6]–[9]. In this literature, it was shown that a controller ensuring robust stability of the closed system with an H_∞ norm restriction can be obtained by solutions of the Riccati inequalities. The background of these results is that the robust stability of the system with parameter perturbation can be guaranteed by a positive definite solution of the Riccati equation (see [10]–[12]). The state feedback problem ([6], [8]) and the measurement feedback ([7], [9]) have been solved using the state-space approach.

On the other hand, the state-space approach was extended to nonlinear H_∞ control problems [2] and [3] where the terminology of nonlinear H_∞ control means that the L_2 induced norm from input to output is less than one. This extension was possible based on the relation between the L_2 induced norm of nonlinear system and the Hamilton–Jacobi equation or inequality. The relation is a generalization of the one between H_∞ norm and Riccati equation in linear systems. The classical notion of dissipativeness in [4] and [5] plays an important role to obtain this result. A nonlinear state feedback [2] and measurement feedback [3] approaches were developed, respectively, along this research line.

In this note, the line of the research is kept to develop robust H_∞ suboptimal control of uncertain nonlinear systems. We will extend the approach given in [6] to solve the robust H_∞ performance problem for nonlinear systems with gain bounded uncertainty. In other words, a state feedback controller such that the L_2 induced norm of the closed-loop system with the uncertainty is less than one can be

obtained by a smooth solution of Hamilton–Jacobi inequality. The approach developed in this note can be extended to output feedback case.

The following notations are adopted here. For vector $z \in R^n$, $\|z\|^2$ and $\|z\|_T$ denote $z^T z$ and $\{\int_0^T z^T(\tau)z(\tau)d\tau\}^{1/2}$, respectively. For a differentiable function $V: R^n \rightarrow R$, $\frac{\partial V}{\partial x}(x)$ is the row-vector of partial derivatives. $L_2(0, T)$ denotes a set of vector-valued function $z(t)$ satisfying $\|z(t)\|_T < \infty$.

II. PRELIMINARIES

Consider a nonlinear system with a gain bounded perturbation given by

$$\dot{x} = f(x) + \Delta f(x) + g_1(x)d + g_2(x)u \quad (1)$$

$$z = h_1(x) + k_{12}(x)u \quad (2)$$

where x is a state vector defined on a neighborhood \mathcal{X} of the origin in R^n , $u \in R^m$ denotes the control input, $d \in R^r$ the disturbance input, $z \in R^p$ the penalty variable, f, g_1, g_2, h_1 and k_{12} are known smooth mappings, and Δf is a smooth uncertain mapping. It is assumed that $f(0) = 0$ and $h_1(0) = 0$ without loss of generality.

We make the following assumption throughout this note.

Assumption 1: The uncertain mapping $\Delta f(x)$ is described by $\Delta f(x) = c(x)\delta(x)$ with known smooth mapping $c: R^n \rightarrow R^{n \times q_1}$ and unknown smooth mapping $\delta: R^n \rightarrow R^{q_1}$. Δf belongs to a bounded set defined by

$$\Omega := \{\Delta f(x) \mid \delta(0) = 0, \quad \|\delta(x)\|^2 \leq \|w(x)\|^2, \forall x \in \mathcal{X}\} \quad (3)$$

where $w: R^n \rightarrow R^{q_2}$ is a given weighting mapping.

Assumption 2: $k_{12}^T(x)[h_1(x) \ k_{12}(x)] = [0 \ I]$.

Assumption 2 is a nonlinear version of the standard assumption in [1] for simplicity.

For the system (1)–(2), consider a nonlinear state-feedback law

$$u = \alpha(x), \quad \alpha(0) = 0. \quad (4)$$

Definition 1: Let $\gamma > 0$ be given. The closed-loop system of (1) and (2) with (4) is said to have locally robust disturbance attenuation performance in $\mathcal{U} \subset \mathcal{X}$ if for any $\Delta f \in \Omega$, the free system with $d = 0$ is locally asymptotically stable with domain of attraction containing \mathcal{U} , and $\|z\|_T \leq \gamma\|d\|_T$ for every $T > 0$ and every $d \in \mathcal{D}$, where the set \mathcal{D} is defined by

$$\mathcal{D} := \{d \mid d \in L_2(0, T) \text{ s.t. } x(t) \text{ with } x(0) = 0 \text{ remains in } \mathcal{U}, \forall t \leq T\}. \quad (5)$$

The nonlinear robust H_∞ control problem is now defined as follows. Given the plant (1)–(2), find a state feedback controller (4) such that the closed-loop system has local robust disturbance attenuation performance. Without loss of generality, let $\gamma = 1$.

We collect some preliminary results in the next Lemma (see [2] and [4] for proof) which are useful to prove our main result.

Consider a smooth nonlinear system given by

$$\dot{x} = f(x) + g(x)d \quad (6)$$

$$z = h(x) \quad (7)$$

where the state x is defined on \mathcal{X} , $d \in R^r$ and $z \in R^p$ denote the input and the output, respectively. f, g , and h are smooth mappings satisfying $f(0) = 0$ and $h(0) = 0$.

Manuscript received January 21, 1994; revised June 2, 1994.

The authors are with the Department of Mechanical Engineering, Sophia University, Kioicho 7-1, Chiyoda-ku, Tokyo, 102 Japan.

IEEE Log Number 9408793.

Lemma 1: Assume that the system (6)–(7) is zero-state observable [2]. The free system of (6) with $d = 0$ is asymptotically stable and $\|z\|_T \leq \|d\|_T$ for every $T > 0$, if system (6)–(7) is dissipative with respect to the supply rate [5]

$$w(d, z) = \frac{1}{2} \{ \|d\|^2 - \|z\|^2 \} \quad (8)$$

or equivalently, there exists a nonnegative definite smooth solution $V(x) \geq 0$ satisfying the Hamilton-Jacobi inequality

$$\begin{aligned} \frac{\partial V}{\partial x}(x)f(x) + \frac{1}{2} \frac{\partial V}{\partial x}(x)g(x)g^T(x) \frac{\partial^T V}{\partial x}(x) \\ + \frac{1}{2} h^T(x)h(x) \leq 0, \quad V(0) = 0. \end{aligned} \quad (9)$$

III. MAIN RESULT

The main result is to give a solution of the robust H_∞ control problem. We start with finding a sufficient condition for the nonlinear system having robust H_∞ performance.

Consider a nonlinear system with uncertainty described by

$$\dot{x} = f(x) + \Delta f(x) + g(x)d \quad (10)$$

$$z = h(x) \quad (11)$$

where the uncertain nonlinear mapping $\Delta f(x)$ satisfies Assumption 1.

For system (10)–(11), define an extended nonlinear system by

$$\begin{aligned} \dot{x} &= f(x) + [g(x)\lambda(x)e(x)]d' \\ h(x) & \end{aligned} \quad (12)$$

where $d' \in R^{r+q}$, $z' \in R^{r+q}$, and $\lambda(x) > 0$ is a given smooth scale function in $x \in \mathcal{X}$.

Theorem 1: Suppose that the system (10)–(11) is zero-state observable for every $\Delta f(x) \in \Omega$. If there exists a smooth scale function $\lambda(x) > 0$ such that the system (12)–(13) with respect to the supply rate $\frac{1}{2} \{ \|d'\|^2 - \|z'\|^2 \}$ is dissipative, then the nonlinear system (10)–(11) has locally robust disturbance attenuation performance in \mathcal{U} .

Proof: Suppose that system (12)–(13) with respect to supply rate $\frac{1}{2} \{ \|d'\|^2 - \|z'\|^2 \}$ is dissipative, then there exists a nonnegative definite smooth function $V(x) \geq 0$ ($V(0) = 0$) such that

$$V(x) \leq \int_0^t \frac{1}{2} \{ \|d'\|^2 - \|z'\|^2 \} d\tau + V(x_0). \quad (14)$$

So that

$$\frac{\partial V}{\partial x}(x) \{ f(x) + [g(x)\lambda(x)e(x)]d' \} \leq \frac{1}{2} \{ \|d'\|^2 - \|z'\|^2 \}. \quad (15)$$

From Theorem 2 in [2], there exists a $V(x) \geq 0$ ($V(0) = 0$) such that

$$\begin{aligned} \frac{\partial V}{\partial x}(x)f(x) + \frac{1}{2} \frac{\partial V}{\partial x}(x)g(x)g^T(x) \frac{\partial^T V}{\partial x}(x) \\ + \lambda^2(x)e(x)e^T(x) \frac{\partial^T V}{\partial x}(x) \\ + \frac{1}{2} \left\{ h^T(x)h(x) + \frac{1}{\lambda^2(x)} w^T(x)w(x) \right\} \leq 0. \end{aligned} \quad (16)$$

Adding and subtracting $\frac{\partial V}{\partial x}(x)\Delta f(x)$ in (16)

$$\begin{aligned} \frac{\partial V}{\partial x}(x) \{ f(x) + \Delta f(x) \} \\ + \frac{1}{2} \frac{\partial V}{\partial x}(x)g(x)g^T(x) \frac{\partial^T V}{\partial x}(x) + \frac{1}{2} h^T(x)h(x) \\ - \frac{\partial V}{\partial x}(x)v(x)\delta(x) \\ + \frac{\lambda^2(x)}{2} \frac{\partial V}{\partial x}(x)e(x)e^T(x) \frac{\partial^T V}{\partial x}(x) \\ + \frac{1}{2\lambda^2(x)} w^T(x)w(x) \leq 0. \end{aligned} \quad (17)$$

Rearranging some terms in (17), we have

$$\begin{aligned} \frac{\partial V}{\partial x}(x) \{ f(x) + \Delta f(x) \} \\ + \frac{1}{2} \frac{\partial V}{\partial x}(x)g(x)g^T(x) \frac{\partial^T V}{\partial x}(x) + \frac{1}{2} h^T(x)h(x) \\ + \frac{\lambda^2(x)}{2} \left\| e^T(x) \frac{\partial^T V}{\partial x}(x) - \frac{1}{\lambda^2(x)} \delta(x) \right\|^2 \\ + \frac{1}{2\lambda^2(x)} \{ \|w(x)\|^2 - \|\delta(x)\|^2 \} \leq 0. \end{aligned} \quad (18)$$

Since $\|\delta(x)\|^2 \leq \|w(x)\|^2$ for every $x \in \mathcal{X}$ and every $\Delta f(x) \in \Omega$, we obtain

$$\begin{aligned} \frac{\partial V}{\partial x}(x) \{ f(x) + \Delta f(x) \} \\ + \frac{1}{2} \frac{\partial V}{\partial x}(x)g(x)g^T(x) \frac{\partial^T V}{\partial x}(x) \\ + \frac{1}{2} h^T(x)h(x) \leq 0, \quad \Delta f \in \Omega. \end{aligned} \quad (19)$$

Therefore, by Lemma 1 the free system (10)–(11) with $d = 0$ is asymptotically stable and $\|z\|_T \leq \|d\|_T, \forall d \in D$ for every $\Delta f(x) \in \Omega$.

In consequence of Theorem 1, we have the following corollary.

Corollary 1. Suppose that the system (10)–(11) is zero-state observable for every $\Delta f \in \Omega$. The system has the locally robust disturbance attenuation performance in \mathcal{U} if there exists a smooth scale function $\lambda(x) > 0$ such that the Hamilton-Jacobi inequality

$$\begin{aligned} \frac{\partial V}{\partial x}(x)f(x) + \frac{1}{2} \frac{\partial V}{\partial x}(x)g(x)g^T(x) \\ + \lambda^2(x)e(x)e^T(x) \frac{\partial^T V}{\partial x}(x) \\ + \frac{1}{2} \left\{ h^T(x)h(x) + \frac{1}{\lambda^2(x)} w^T(x)w(x) \right\} \leq 0 \end{aligned} \quad (20)$$

has smooth solution $V(x) \geq 0$ ($V(0) = 0$).

Remark 1: It is interesting to compare the result in Corollary 1 with linear case given in [6] and [8]. Consider a linear system with parameter perturbation

$$\dot{x} = (A + E\Sigma(t))x + Bd \quad (21)$$

$$z = Cx \quad (22)$$

where unknown time-varying matrix $\Sigma(t)$ satisfies $\Sigma^T(t)\Sigma(t) \leq F^T F, \forall t$ for a given F . Since a smooth solution $V(x)$ of the Hamilton-Jacobi inequality can be given by $V(x) = x^T P x$, where P is the nonnegative definite solution of the Riccati inequality

$$A^T P + P A + P(BB^T + \lambda^2 E E^T)P + C^T C + \frac{1}{\lambda^2} F^T F < 0 \quad (23)$$

for a scale $\lambda > 0$, Corollary 1 implies that the linear system (21)–(22) has the robust H_∞ performance if the Riccati inequality has the nonnegative definite solution, which has been shown in [6].

Remark 2: Note that in nonlinear case the scaling parameter λ be chosen as a function of x , which may reduce conservativeness of the result.

We now consider the robust H_∞ control problem mentioned in Section II. Using the condition in Theorem 1, the problem can be solved by finding such a state feedback controller that makes the

closed-loop system dissipative with respect to the supply rate. It can be done by solving an Hamilton-Jacobi inequality.

Theorem 2 Assume that the system $\dot{x} = f(x) + \Delta f(x) + g_1(x)d$, $z = h_1(x)$ is zero-state observable for every $\Delta f(x) \in \Omega$. If there exists a scale smooth function $\lambda(x) > 0$ such that the Hamilton-Jacobi inequality

$$\begin{aligned} & \frac{\partial V}{\partial x}(x)f(x) + \frac{1}{2} \frac{\partial V}{\partial x}(x) \\ & \times \left\{ g_1(x)g_1^T(x) + \lambda^2(x)x(x)x^T(x) - g_2(x)g_2^T(x) \right\} \frac{\partial^T V}{\partial x}(x) \\ & + \frac{1}{2} \left\{ h_1^T(x)h_1(x) + \frac{1}{\lambda^2(x)}u^T(x)u(x) \right\} \leq 0 \end{aligned} \quad (24)$$

has a smooth solution $V(x) \geq 0$ ($V(0) = 0$), then a solution of the robust H_∞ control problem is given by

$$\alpha(x) = -g_2^T(x) \frac{\partial^T V}{\partial x}(x) \quad (25)$$

Proof Note that the closed loop system of the plant (1)-(2) with the state feedback controller $u = \alpha(x)$ is given by

$$\dot{x} = f(x) + \Delta f(x) + g_2(x)\alpha(x) + g_1(x)d \quad (26)$$

$$z = h_1(x) + k_{12}(x)\alpha(x) \quad (27)$$

Using the Hamilton-Jacobi inequality (24) and the assumptions in Section II, we obtain

$$\begin{aligned} & \frac{\partial V}{\partial x}(x)\{f(x) + \Delta f(x) + g_2(x)\alpha(x)\} \\ & + \frac{1}{2} \frac{\partial V}{\partial x}(x)g_1(x)g_1^T(x) \frac{\partial^T V}{\partial x}(x) \\ & + \frac{1}{2} \left\{ h_1^T(x) + \alpha^T(x)k_{12}^T(x) \right\} \left\{ h_1(x) + k_{12}(x)\alpha(x) \right\} \\ & \leq -\frac{1}{2} \left\| \lambda(x)x^T(x) \frac{\partial^T V}{\partial x}(x) - \frac{1}{\lambda(x)}\delta(x) \right. \\ & \quad \left. - 2\lambda^2(x) \frac{\partial^T V}{\partial x}(x) \right\|^2 - \|\delta(x)\|^2 \\ & + \frac{1}{2} g_2^T(x) \frac{\partial^T V}{\partial x}(x) + \alpha(x) \end{aligned} \quad (28)$$

Use of (25) for $\alpha(x)$ and assumption on the uncertain mapping $\delta(x)$ gives

$$\begin{aligned} & \frac{\partial V}{\partial x}(x)\{f(x) + \Delta f(x) + g_2(x)\alpha(x)\} \\ & + \frac{1}{2} \frac{\partial V}{\partial x}(x)g_1(x)g_1^T(x) \frac{\partial^T V}{\partial x}(x) \\ & + \frac{1}{2} \left\{ h_1^T(x) + \alpha^T(x)k_{12}^T(x) \right\} \left\{ h_1(x) + k_{12}(x)\alpha(x) \right\} \\ & \leq 0, \quad \forall \Delta f(x) \in \Omega \end{aligned} \quad (29)$$

By Theorem 2 in [2]

$$\begin{aligned} & \frac{\partial V}{\partial x}(x)\{f(x) + \Delta f(x) + g_2(x)\alpha(x)\} + \frac{\partial V}{\partial x}(x)g_1(x)d \\ & \leq \frac{1}{2}(\|d\|^2 - \|z\|^2), \quad \forall \Delta f(x) \in \Omega \end{aligned} \quad (30)$$

Hence, for every $T > 0$, $\|z\|_T \leq \|d\|_T$, $\forall d \in \mathcal{D}$ for every $\Delta f(x) \in \Omega$

To show the robust asymptotic stability of the free system

$$\dot{x} = f(x) + g_2(x)\alpha(x) + \Delta f(x) \quad (31)$$

we consider the solution $V(x) \geq 0$ as a Lyapunov function. From (29) and Assumption 2, we have $V \leq 0$, $\forall \Delta f(x) \in \Omega$ and $V = 0$ implies $h_1(x) = 0$ and $\alpha(x) = 0$. Therefore, the robust asymptotic stability follows from the Assumption that pair $\{f + \Delta f, h\}$ is zero state observable for every Δf .

Remark 3 As shown in Remark 1 in linear case the solution $V(x) \geq 0$ of (24) will be given by the solution of the Riccati inequality (see the result in [6]).

IV. CONCLUSION

In this note, a sufficient condition for that a nonlinear system with uncertainty has robust H_∞ performance is given based on the classical dissipativeness theory. Using this condition such a state feedback control law is derived based on the Hamilton-Jacobi inequality that the closed loop system has locally robust disturbance attenuation performance. It is also shown that the well known results for linear system are special cases of the presented results.

ACKNOWLEDGMENT

The authors thank Prof. A. J. van der Schaft and anonymous referees for some helpful comments on an earlier version of the manuscript.

REFERENCES

- [1] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, "State space solution to standard H_2 and H_∞ control problems," *IEEE Trans Automat Contr*, vol. 34, pp. 831-846, 1989.
- [2] A. J. van der Schaft, "L₂ gain analysis of nonlinear systems and nonlinear state feedback H_∞ control," *IEEE Trans Automat Contr*, vol. 37, pp. 770-783, 1992.
- [3] A. Isidori and A. Astolfi, "Disturbance attenuation and H_∞ control via measurement feedback in nonlinear systems," *IEEE Trans Automat Contr*, vol. 37, pp. 1283-1293, 1992.
- [4] D. J. Hill and P. J. Moylan, "The stability of nonlinear dissipative systems," *IEEE Trans Automat Contr*, vol. AC-21, pp. 708-711, 1976.
- [5] J. C. Willems, "Dissipative dynamical systems. part I: general theory," *Arch Rat Mech Anal*, vol. 45, pp. 321-351, 1972.
- [6] L. Xie and C. de Souza, "Robust H_∞ control for linear systems with norm bounded time varying uncertainty," *IEEE Trans Automat Contr*, vol. 37, pp. 1188-1191, 1992.
- [7] L. Xie, M. Fu, and C. F. de Souza, " H_∞ control and quadratic stabilization of systems with parameter uncertainty via output feedback," *IEEE Trans Automat Contr*, vol. 37, pp. 1253-1256, 1992.
- [8] I. Shen and K. Tamura, "Properties of algebraic Riccati inequalities and H_∞ robust suboptimal controller design," in *Proc IEEE 30th CDC Brighton*, vol. 3, 1991, pp. 212-213.
- [9] —, "State feedback performance recovery via an observer in H_∞ robust control," *Syst Contr Lett*, vol. 21, pp. 463-474, 1993.
- [10] P. P. Khargonekar, I. R. Petersen, and K. Zhou, "Robust stabilization of uncertain linear systems: quadratic stabilizability and H_∞ control theory," *IEEE Trans Automat Contr*, vol. 35, pp. 356-361, 1990.
- [11] D. Hinrichsen and A. J. Pritchard, "Stability radius for structured perturbations and the algebraic Riccati equation," *Syst Contr Lett*, vol. 8, pp. 1055-1113, 1986.
- [12] I. R. Petersen and C. V. Hollot, "A Riccati equation approach to stabilization of uncertain linear systems," *Automatica*, vol. 22, pp. 397-411, 1986.

Optimal Nonparametric Identification from Arbitrary Corrupt Finite Time Series

Jie Chen, Carl N. Nett, and Michael K. H. Fan

Abstract—In this paper we formulate and solve a worst-case system identification problem for single-input, single-output, linear, shift-invariant, distributed parameter plants. The available *a priori* information in this problem consists of time-dependent upper and lower bounds on the plant impulse response and the additive output noise. The available *a posteriori* information consists of a corrupt finite output time series obtained in response to a known, nonzero, but otherwise arbitrary, input signal. We present a novel identification method for this problem. This method maps the available *a priori* and *a posteriori* information into an “uncertain model” of the plant, which is comprised of a nominal plant model, a bounded additive output noise, and a bounded additive model uncertainty. The upper bound on the model uncertainty is explicit and expressed in terms of both the ℓ_1 and H_∞ system norms. The identification method and the nominal model possess certain well-defined optimality properties and are computationally simple, requiring only the solution of a single linear programming problem.

I. INTRODUCTION

The primary goal of this paper is to develop a system identification method for identifying, from available *a priori* and *a posteriori* information, an “uncertain plant model” which can be used for robust control design. Typically, an uncertain model consists of a nominal design model, a noise specification, and a model error specification. Though the specific nature of the required noise and model error specifications is dictated by the specific robust control method to be applied, nearly all existing robust control design methods require that these specifications be stated as explicit, worst-case/deterministic bounds on the levels of noise and model error [7], [8]. For this reason a worst-case/deterministic identification approach is adopted in this paper. System identification methods that take a worst-case/deterministic, robust control oriented approach have recently received considerable interest; an extensive list of publications on these problems may be found in, e.g., [12] and [36] and the references therein.

The specific identification problem considered in this paper pertains to single-input, single-output, linear, shift-invariant, distributed parameter plants. The available *a priori* information in this problem consists of time-dependent upper and lower bounds on the plant impulse response and the additive output noise. The available *a posteriori* information consists of a corrupt finite output time series obtained in response to a known, nonzero, but otherwise arbitrary applied input. Our objective is to develop from the given information an uncertain model which has an additive model error structure and whose modeling uncertainty is measured in the ℓ_1 and H_∞ system norms. Our motivation for these choices of model error structures and system norms stems from the fact that they are consistent with the currently popular ℓ_1 and H_∞ design methods [7], [8]. Recently,

there has been a bulk of research work devoted to problems of this kind, under the name of identification in ℓ_1 and H_∞ . We refer the reader to some of the closely related problems and results in, e.g., [12], [15], [10], [27], [4], [21], [20], [33], [5], [16], [1], and [29] for a glimpse of the work in this area; discussions on these results are beyond the scope of this note.

Our main contribution in this paper consists of an identification algorithm and several model error bounds. The identification algorithm maps the given *a priori* and *a posteriori* information to an identified nominal model. The error bounds are explicit and are expressed in terms of both the ℓ_1 and H_∞ norms. Our algorithm possesses certain well-defined optimality properties and is generally referred to as an interpolatory algorithm in the information-based complexity (IBC) theory [31], [32]. It is a standard result in IBC that algorithms of the interpolatory class are worst-case strongly optimal to within a factor of two, in the sense that the worst-case error achievable in applying the algorithm to approximate the underlying plant is at most a factor of two of the smallest possible such error. Moreover, our algorithm also has the advantage that it can be constructed in an extremely simple manner: it requires only that a single linear programming problem be solved. Technically, our development shares some elements with the set membership method for parametric identification problems (see, e.g., [23]–[25] and the references therein).

The notation used throughout this paper is as follows. Let \mathbf{Z} denote the set of integers, $\mathbf{Z}_+ := \{k \in \mathbf{Z} : k \geq 0\}$, and $\mathbf{Z}_{+,n} := \{k \in \mathbf{Z}_+ : 0 \leq k \leq n-1\}$. Let \mathbf{R} denote the set of real numbers, and \mathbf{R}^n denote the space of n dimensional real vectors. Let $S_n(a, b) := \{x \in \mathbf{R}^n : a_k \leq x_k \leq b_k, k \in \mathbf{Z}_{+,n}\}$. Define the normed spaces $\ell_1 := \{f: \mathbf{Z}_+ \rightarrow \mathbf{R} \mid \|f\|_1 := \sum_{k=0}^{\infty} |f(k)| < \infty\}$, $\ell_2 := \{f: \mathbf{Z}_+ \rightarrow \mathbf{R} \mid \|f\|_2^2 := \sum_{k=0}^{\infty} |f(k)|^2 < \infty\}$, and $\ell_\infty := \{f: \mathbf{Z}_+ \rightarrow \mathbf{R} \mid \|f\|_\infty := \sup_{k \in \mathbf{Z}_+} |f(k)| < \infty\}$. Let $\mathbf{D} := \{z \in \mathbf{C} : |z| < 1\}$ and denote $H_\infty := \{f: \mathbf{D} \rightarrow \mathbf{C} \mid f \text{ analytic in } \mathbf{D} \text{ and } \|f\|_{H_\infty} := \sup_{z \in \mathbf{D}} |f(z)| < \infty\}$. Additionally, define the operator $T_n: \ell_\infty \rightarrow \mathbf{R}^n$ such that $(T_n f)(k) = f(k)$ for $k \in \mathbf{Z}_{+,n}$.

Finally, a preliminary version of this paper was presented previously in [3].

II. PROBLEM FORMULATION AND PRELIMINARY RESULTS

We consider the class of causal, single-input, single-output, linear, shift-invariant, stable distributed parameter systems. This class of systems can be identified with a normed linear space $\mathbf{X} \subset \ell_1$ which contains one-sided sequences and is endowed with a system induced norm $\|\cdot\|$. A system $h \in \mathbf{X}$ is represented by its impulse response $\{h(k)\}_{k=0}^{\infty}$, or alternatively, by the corresponding transform $\hat{h}(z)$, where¹

$$\hat{h}(z) := \sum h(k)z^k. \quad (2.1)$$

The system to be identified is assumed to belong to \mathbf{X} . Additionally, its impulse response $\{h(k)\}_{k=0}^{\infty}$ is assumed to satisfy $M_k^- \leq h(k) \leq M_k^+, \forall k \in \mathbf{Z}_+$, where $\{M_k^-\}_{k=0}^{\infty} \in \ell_1$ and $\{M_k^+\}_{k=0}^{\infty} \in \ell_1$ are two prespecified sequences which represent the available plant *a priori* information. We may write $h \in S_\infty(M^-, M^+) \subset \mathbf{X}$.

The experimental procedure considered consists of applying an arbitrary nonzero input $u \in \ell_\infty$ to the system $h \in \mathbf{X}$ to generate an

¹This transform corresponds to the standard z -transform evaluated at $1/z$.

Manuscript received January 12, 1994; revised June 3, 1994. This work was supported in part by NSF and ONR.

J. Chen is with the College of Engineering, University of California, Riverside, CA 92521-0425 USA.

C. N. Nett is with United Technology Research Center, East Hartford, CT 06108 USA.

M. K. Fan is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250 USA.

IEEE Log Number 9408792.

output $h * u$. An additively corrupted version of $h * u$ is observed over a finite duration n . Let the corrupting noise be denoted by $v \in \ell_\infty$. Then, the experimental procedure can be precisely written as

$$E_n(h, v) := T_n((h * u) + v).$$

The *a priori* information on the output noise $v \in \ell_\infty$ consists of two prespecified sequences $\{\epsilon_k^-\}_{k=0}^\infty \in \ell_\infty$ and $\{\epsilon_k^+\}_{k=0}^\infty \in \ell_\infty$. The noise sequence $\{v(k)\}_{k=0}^\infty$ is assumed to satisfy $\epsilon_k^- \leq v(k) \leq \epsilon_k^+$, $\forall k \in \mathbb{Z}_+$ or, alternatively, $v \in S_\infty(\epsilon^-, \epsilon^+) \subset \ell_\infty$. The experiment operator $E_n(\cdot, \cdot) : \mathbf{X} \times \ell_\infty \rightarrow \mathbf{R}^n$ operates on the ordered pair (h, v) and yields an output data record $y \in \mathbf{R}^n$, $y(k) := (E_n(h, v))_k$ for all $k \in \mathbb{Z}_{+n}$. Let U be the lower triangular Toeplitz matrix formed by the input u

$$U := \begin{pmatrix} u(0) & 0 & 0 \\ u(1) & u(0) & 0 \\ & & \vdots \\ u(n-1) & u(n-2) & u(0) \end{pmatrix}$$

Then, the output data $y \in \mathbf{R}^n$ can be conveniently expressed as $y = UT_n h + T_n v$, where $h \in S_\infty(M^-, M^+)$, and $v \in S_\infty(\epsilon^-, \epsilon^+)$. The set of all possible such data is given by

$$\mathbf{Y} := \{y \in \mathbf{R}^n : y = UT_n h + T_n v \text{ for some } h \in S_\infty(M^-, M^+), v \in S_\infty(\epsilon^-, \epsilon^+)\}.$$

For each $y \in \mathbf{Y}$, we define the set of indistinguishable systems as

$$\mathbf{P}(y) := \{h \in S_\infty(M^-, M^+) : y = UT_n h + T_n v \text{ for some } v \in S_\infty(\epsilon^-, \epsilon^+)\}.$$

This set contains all plant models which are "unfalsified" and whose members cannot be rigorously distinguished by the available *a priori* and *a posteriori* information. Let $\mathbf{P}_n(y) := \{T_n h : h \in \mathbf{P}(y)\}$. Then, $\mathbf{P}_n(y)$ consists of the truncated series in $\mathbf{P}(y)$. One may readily observe that $\mathbf{P}_n(y)$ is a convex polytope in \mathbf{R}^n , as it is defined via a finite set of linear inequalities.

Given the above *a priori* and experimental information, our primary goal in this paper is to obtain an uncertain model through the development of an identification algorithm and the quantification of a corresponding identification error. An identification algorithm is a mapping² $A_n : \mathbf{R}^n \rightarrow \mathbf{X}$ which operates on the data record and generates a model $A_n(y) \in \mathbf{X}$. The identification error associated with an algorithm A_n is measured by the norm $\|\cdot\|$ and is used as a bound on the modeling uncertainty. Several notions concerning identification algorithms and errors are adapted below from the existing IBC and identification literature (cf. [31], [23], [24], [12]):

- local identification error: $e(A_n; y; \|\cdot\|) := \sup_{h \in \mathbf{P}(y)} \|h - A_n(y)\|$,
- global identification error: $e(A_n; \|\cdot\|) := \sup_{y \in \mathbf{Y}} e(A_n; y; \|\cdot\|)$,
- optimal local identification error: $e^*(y; \|\cdot\|) := \inf_{A_n \in \mathcal{A}_n(y) \in \mathbf{X}} e(A_n; y; \|\cdot\|)$,
- optimal global identification error: $e^*(\|\cdot\|) := \inf_{A_n \in \mathcal{A}_n(y) \in \mathbf{X}} e(A_n; \|\cdot\|)$.

An algorithm A_n^* is said to be locally optimal at y if $e(A_n^*; y; \|\cdot\|) = e^*(y; \|\cdot\|)$. If it is locally optimal at each $y \in \mathbf{Y}$, then it is said to be strongly optimal.

²In a broader sense, an identification algorithm acts on the input-output data pair (u, y) , as the input signal plays an important role in identification. It is not our aim, however, to discuss the role of input signals in this note. Hence, to simplify the notation, we omit the dependence of all the relevant notions (e.g., $\mathbf{P}(y)$, $A_n(y)$, etc.) on the input and the *a priori* information.

In the IBC framework, the optimal algorithm and error can be alternatively characterized as the center and radius of the set $\mathbf{P}(y)$. For a given data record $y \in \mathbf{Y}$, let the radius and diameter [31], [13] of $\mathbf{P}(y)$ be denoted by $r(\mathbf{P}(y); \|\cdot\|)$ and $d(\mathbf{P}(y); \|\cdot\|)$, respectively. Then, the following result is known [31].

Fact 2.1: For any data record y , $e^*(y; \|\cdot\|) = r(\mathbf{P}(y); \|\cdot\|)$ and $d(\mathbf{P}(y); \|\cdot\|)/2 \leq e^*(y; \|\cdot\|) \leq d(\mathbf{P}(y); \|\cdot\|)$. Furthermore, for any $p \in \mathbf{P}(y)$, $r(\mathbf{P}(y); \|\cdot\|) \leq \sup_{h \in \mathbf{P}(y)} \|h - p\| \leq d(\mathbf{P}(y); \|\cdot\|)$.

As in [31], we shall call $r(\mathbf{P}(y); \|\cdot\|)$ and $d(\mathbf{P}(y); \|\cdot\|)$ the local radius and diameter of information. Accordingly, the global radius and diameter of information are designated to the quantities $\sup_{y \in \mathbf{Y}} r(\mathbf{P}(y); \|\cdot\|)$ and $\sup_{y \in \mathbf{Y}} d(\mathbf{P}(y); \|\cdot\|)$. The center of $\mathbf{P}(y)$ is the element $p \in \mathbf{X}$ such that $r(\mathbf{P}(y); \|\cdot\|) = \sup_{h \in \mathbf{P}(y)} \|h - p\|$. If for some $A_n, p = A_n(y)$ for all $y \in \mathbf{Y}$, then A_n is called the central algorithm [31], [23]. Clearly, an algorithm is strongly optimal if and only if it is central. In general, it is difficult to find the center of $\mathbf{P}(y)$. If, however, $\mathbf{P}(y)$ is a symmetric set, then its center can be easily found. The set $\mathbf{P}(y)$ is said to be symmetric if there exists a $p \in \mathbf{X}$ such that for any $h \in \mathbf{P}(y)$, $p + h \in \mathbf{P}(y)$ implies $p - h \in \mathbf{P}(y)$. The element p possessing this property is called the point of symmetry in $\mathbf{P}(y)$. It was shown in [31] that a point of symmetry is a center, and hence it defines a central algorithm. We shall say that such a central algorithm is symmetrically central. It is useful to point out that a symmetrically central algorithm is central for all norms, as a point of symmetry is of a purely geometric nature.

In this paper, we are interested in a class of nearly optimal algorithms termed interpolatory algorithms [31]. An algorithm $A_n : \mathbf{R}^n \rightarrow \mathbf{X}$ is said to be interpolatory if $A_n(y) \in \mathbf{P}(y)$. An appealing property of interpolatory algorithms is that they yield an identification error that is tight to within a factor of two of the optimal error for all possible data records [31]. This is clear from Fact 2.1 and is summarized in the following.

Fact 2.2: Let A_n be an algorithm such that $A_n(y) \in \mathbf{P}(y)$. Then, $e^*(y; \|\cdot\|) \leq e(A_n; y; \|\cdot\|) \leq 2e^*(y; \|\cdot\|)$.

In the sense of Fact 2.2, we say that an interpolatory algorithm is strongly optimal to within a factor of two. It is clear that this property holds regardless of the norms. Note also from Fact 2.1 that the identification error for any interpolatory algorithm is bounded from above by $d(\mathbf{P}(y); \|\cdot\|)$ and from below by $d(\mathbf{P}(y); \|\cdot\|)/2$. This fact makes it possible to estimate the error by computing the diameter of information. The following characterization [31] may be used to compute the global diameter $\sup_{y \in \mathbf{Y}} d(\mathbf{P}_n(y); \|\cdot\|)$.

Fact 2.3: Let $M_k^- = -M_k^+$ and $\epsilon_k^- = -\epsilon_k^+$ for all $k \in \mathbb{Z}_{+n}$. Consider the normed space $(\mathbf{R}^n, \|\cdot\|)$. Then

$$\sup_{y \in \mathbf{Y}} d(\mathbf{P}_n(y); \|\cdot\|) = 2 \sup_{y \in \mathbf{P}_n(0)} \|\cdot\| \quad (2.2)$$

The key condition implicit in Fact 2.3 is that the hyperrectangle $S_n(M^-, M^+) \times S_n(\epsilon^-, \epsilon^+) \subset \mathbf{R}^{2n}$ is a balanced convex set [31]. An extension of this result will be given in the next section.

III. MAIN RESULTS

In this section, we shall first formulate and solve a problem concerning the consistency of the *a priori* and *a posteriori* information. This problem bears some similarity to the model validation problems studied in [30] and [28]. Significant differences also exist, however, between the two problems. Based on the solution of this problem, we construct an interpolatory algorithm and derive bounds for the corresponding identification error. Finally, we provide a brief discussion on central algorithms and optimal identification errors.

A. Consistency of Information

The problem of consistency between the *a priori* information and data record is concerned with the following question: Given the model set defined by $S_\infty(M^-, M^+)$, the noise set $S_\infty(\epsilon^-, \epsilon^+)$, and given also a data record $y \in \mathbf{R}^n$, will there exist an $h \in S_\infty(M^-, M^+)$ so that $y = T_n((h * u) + v)$? This problem can be equivalently posed as to determining whether the set $\mathbf{P}(y)$ is empty.

Definition 3.1: A data record $y \in \mathbf{R}^n$ is said to be consistent with the plant and noise *a priori* information if $\mathbf{P}(y) \neq \emptyset$.

It is important to note that the data and *a priori* information should be consistent for the identification problem to be meaningful, for if $\mathbf{P}(y)$ is empty, the identification error will become unbounded.

It can be readily recognized that $\mathbf{P}(y) \neq \emptyset$ if and only if $\mathbf{P}_n(y) \neq \emptyset$. This observation leads to the following simple necessary and sufficient condition that can be used to check the consistency between the data and *a priori* information.

Lemma 3.1: A data record $y \in \mathbf{R}^n$ is consistent with the *a priori* information if and only if the linear inequalities

$$M_k^- \leq h(k) \leq M_k^+, \quad k = 0, 1, \dots, n-1, \quad (3.1)$$

$$\epsilon_k^- \leq y(k) - \sum_{j=0}^k u(k-j)h(j) \leq \epsilon_k^+, \quad k = 0, 1, \dots, n-1 \quad (3.2)$$

admit a solution.

Consequently, the consistency problem amounts to solving a finite set of linear inequalities, which in turn may be solved efficiently using the linear programming method (see, e.g., [26]). It is easy to derive necessary or sufficient conditions by weakening (3.1)–(3.2). It is also possible to extend this result to solve a companion problem which is to compute a minimal noise envelope such that the data and *a priori* information are consistent. The reader is referred to [3] for a detailed discussion on these issues.

B. An Interpolatory Algorithm and Error Bounds

The interpolatory algorithm to be proposed below is motivated from the IBC theory [31] and requires only solving the data consistency problem. More specifically, it follows from Fact 2.2 and Lemma 3.1 that any solution of (3.1) and (3.2) serves as an identified model that yields a local identification error to within a factor of two of the optimal local error. Hence, the following result is clear.

Theorem 3.1: Let $\{h(k)\}_{k=0}^{n-1}$ be a solution to (3.1)–(3.2). Construct A_n so that

$$[A_n(y)]_k = \begin{cases} h(k) & 0 \leq k \leq n-1 \\ \frac{M_k^+ + M_k^-}{2} & k \geq n. \end{cases} \quad (3.3)$$

Then, for any data record $y \in \mathbf{Y}$, $\epsilon^*(y; \|\cdot\|) \leq \epsilon(A_n; y; \|\cdot\|) \leq 2\epsilon^*(y; \|\cdot\|)$.

As a result of this construction, the algorithm A_n is clearly interpolatory, and it can be readily computed by solving a single linear program. As pointed out in the preceding section, this algorithm is strongly optimal to within a factor of two for all norms.

Having constructed the above interpolatory algorithm, in the remainder of this section we shall derive explicit bounds for the local and global identification errors. Denote

$$\delta M_k := M_k^+ - M_k^-, \quad \delta \epsilon_k := \epsilon_k^+ - \epsilon_k^-. \quad (3.4)$$

We shall first present the following preliminary lemmas.

Lemma 3.2: Let $c := [c_0 \ c_1 \ \dots \ c_{n-1}]^T \in \mathbf{R}^n$ be the center of $\mathbf{P}_n(y)$, and let δM_k be defined by (3.4). Then, the algorithm A_n^* , such that

$$[A_n^*(y)]_k = \begin{cases} c_k & 0 \leq k \leq n-1 \\ \frac{M_k^+ + M_k^-}{2} & k \geq n \end{cases}$$

is central. Furthermore

$$\epsilon^*(y; \|\cdot\|_1) = \frac{1}{2} \sum_{k=n}^{\infty} \delta M_k + r(\mathbf{P}_n(y), \|\cdot\|_1). \quad (3.5)$$

Proof: By definition, we have

$$\begin{aligned} r(\mathbf{P}(y), \|\cdot\|_1) &= \inf_{h \in \mathbf{X}} \sup_{p \in \mathbf{P}(y)} \sum_{k=0}^{\infty} |h(k) - p(k)| \\ &= \inf_{h \in \mathbf{R}^n} \sup_{p \in \mathbf{P}_n(y)} \sum_{k=0}^{n-1} |h(k) - p(k)| \\ &\quad + \inf_{h \in \mathbf{X}/\mathbf{R}^n} \sup_{p \in \mathbf{P}(y)/\mathbf{P}_n(y)} \sum_{k=n}^{\infty} |h(k) - p(k)|. \end{aligned}$$

It is easy to see that $\mathbf{P}(y)/\mathbf{P}_n(y)$ is a symmetric set, and its point of symmetry is h^* , where $h^*(k) = (1/2)(M_k^+ + M_k^-)$ for all $k \geq n$. Hence, it follows that

$$\inf_{h \in \mathbf{X}/\mathbf{R}^n} \sup_{p \in \mathbf{P}(y)/\mathbf{P}_n(y)} \sum_{k=n}^{\infty} |h(k) - p(k)| \leq \frac{1}{2} \sum_{k=n}^{\infty} (M_k^+ - M_k^-).$$

The proof is now completed by using the definition of radius and Fact 2.1. \square

Lemma 3.3: Let $A_n(y)$ be given by (3.3). Then

$$\epsilon(A_n; y; \|\cdot\|_1) = \frac{1}{2} \sum_{k=n}^{\infty} \delta M_k + \max_{h \in \mathbf{P}_n(y)} \|T_n h - T_n A_n(y)\|_1. \quad (3.6)$$

Proof: Similar to that of Lemma 3.2. \square

It follows from Lemmas 3.2 and 3.3 that the problem of finding the central algorithm reduces to that of finding the center of $\mathbf{P}_n(y)$, and the problem of finding an interpolatory algorithm simply reduces to that of finding an element of $\mathbf{P}_n(y)$. The latter observation is precisely the underlying idea for the algorithm (3.3), whose construction for $k \in \mathbf{Z}_{+n}$ amounts to finding an element of $\mathbf{P}_n(y)$, and for $k \geq n$ the center of $\mathbf{P}(y)/\mathbf{P}_n(y)$ is selected. It is important to note from these results that the condition $\{\delta M_k\}_{k=0}^{\infty} \in \ell_1$ is necessary so that the identification error is finite. This is clear because the series $\sum_{k=0}^{\infty} \delta M_k$ and $\sum_{k=n}^{\infty} \delta M_k$ have the same convergence behavior (see, e.g., [17]). This condition is always satisfied because $\{M_k\}_{k=0}^{\infty} \in \ell_1$ and $\{M_k^+\}_{k=0}^{\infty} \in \ell_1$.

In what follows we shall develop bounds for the diameter of $\mathbf{P}_n(y)$. According to Lemma 3.3, these quantities can then be used to calculate bounds for the identification error. With no loss of generality, we assume that $u(0) \neq 0$. It follows readily that U is invertible and that U^{-1} is also a lower triangular Toeplitz matrix. We shall write

$$U^{-1} = \begin{bmatrix} a_0 & 0 & \cdots & 0 \\ a_1 & a_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n-1} & a_{n-2} & \cdots & a_0 \end{bmatrix}.$$

Theorem 3.2: Suppose that $A_n(y)$ is given by (3.3). Let $\tilde{h} \in \mathbf{X}$

be defined as $\hat{h}(k) := (M_k^+ + M_k^-)/2$, and $\bar{v} \in \ell_\infty$ be defined as $\bar{v}(k) := (\epsilon_k^+ + \epsilon_k^-)/2$, for all $k \in \mathbb{Z}_+$. Furthermore, denote $\bar{y} := y - UT_n \bar{h} - T_n \bar{v}$

$$h_u(k) := \min \left\{ \sum_{j=0}^k (a_{k-j} \bar{y}(j) + |a_{k-j}| \delta \epsilon_j), \delta M_k \right\}, \quad (3.7)$$

$$h_l(k) := \max \left\{ \sum_{j=0}^k (a_{k-j} \bar{y}(j) - |a_{k-j}| \delta \epsilon_j), -\delta M_k \right\}. \quad (3.8)$$

Then

$$\begin{aligned} \epsilon(A_n; y; \|\cdot\|_{H_\infty}) &\leq \epsilon(A_n; y; \|\cdot\|_1) \\ &\leq \frac{1}{2} \sum_{k=n}^{\infty} \delta M_k + \sum_{k=0}^{n-1} (h_u(k) - h_l(k)). \end{aligned} \quad (3.9)$$

Proof We show that the bound holds for $\epsilon(A_n; y; \|\cdot\|_1)$, as the inequality between $\epsilon(A_n; y; \|\cdot\|_{H_\infty})$ and $\epsilon(A_n; y; \|\cdot\|_1)$ follows from the well-known fact $\|\cdot\|_{H_\infty} \leq \|\cdot\|_1$ [6]. Toward this end, it suffices to prove that $d(\mathbf{P}_n(y), \|\cdot\|_1) \leq \sum_{k=0}^{n-1} (h_u(k) - h_l(k))$ or $\mathbf{P}_n(y) \subset \mathbf{S}_n(h_l + \bar{h}, h_u + \bar{h})$. Note that for any $h \in \mathbf{P}_n(y)$, we have $M_k^- \leq h(k) \leq M_k^+$, and $h(k) = \sum_{j=0}^k a_{k-j} y(j) + \sum_{j=0}^k a_{k-j} v(j)$, where $k \in \mathbb{Z}_+$, and $v \in \mathbf{S}_\infty(\epsilon^-, \epsilon^+)$. The latter equality is equivalent to $\bar{h}(k) - \hat{h}(k) = \sum_{j=0}^k a_{k-j} y(j) + \sum_{j=0}^k a_{k-j} (v(j) - \bar{v}(j))$. Hence, it follows that $-\delta M_k \leq h(k) - \hat{h}(k) \leq \delta M_k$, and that

$$\begin{aligned} \sum_{j=0}^k a_{k-j} \bar{y}(j) - \sum_{j=0}^k |a_{k-j}| \delta \epsilon_j &\leq h(k) - \hat{h}(k) \\ &\leq \sum_{j=0}^k a_{k-j} y(j) + \sum_{j=0}^k |a_{k-j}| \delta \epsilon_j. \end{aligned}$$

This implies that $h - \hat{h} \in \mathbf{S}_n(h_l, h_u)$, and hence that $h \in \mathbf{S}_n(h_l + \bar{h}, h_u + \bar{h})$ and $\mathbf{P}_n(y) \subset \mathbf{S}_n(h_l + \bar{h}, h_u + \bar{h})$. As a result, $d(\mathbf{P}_n(y), \|\cdot\|_1) \leq d(\mathbf{S}_n(h_l + \bar{h}, h_u + \bar{h}), \|\cdot\|_1)$. However

$$\begin{aligned} d(\mathbf{S}_n(h_l + \bar{h}, h_u + \bar{h}), \|\cdot\|_1) &= d(\mathbf{S}_n(h_l, h_u), \|\cdot\|_1) \\ &= \sum_{k=0}^{n-1} (h_u(k) - h_l(k)). \end{aligned}$$

This completes the proof. \square

In addition to the above bound for the local error, we shall also derive below bounds for the global identification error and establish a convergence condition of the proposed interpolatory algorithm. Here an algorithm A_n is said to be convergent [12] if it satisfies the property

$$\lim_{n \rightarrow \infty} \epsilon(A_n; \|\cdot\|) = 0$$

where $\delta \epsilon := \max_k \delta \epsilon_k$. Note that the condition $\{\delta M_k\}_{k=0}^\infty \in \ell_1$ is necessary for any algorithm to be convergent. We shall first need the following extension of Fact 2.3.

Lemma 3.4: Let \bar{h} and \bar{v} be defined as in Theorem 3.2. Then

$$\sup_{y \in \mathbf{Y}} d(\mathbf{P}_n(y), \|\cdot\|) = 2 \sup_{T_n \bar{h} \in \mathbf{P}_n(T_n \bar{h} + T_n \bar{v})} \|T_n \bar{h} - T_n \hat{h}\|. \quad (3.10)$$

Proof: For any $h \in \mathbf{S}_\infty(M^-, M^+)$ and $v \in \mathbf{S}_\infty(\epsilon^-, \epsilon^+)$, we define $\bar{h} := h - \hat{h}$ and $\bar{v} := v - \bar{v}$. Then, $h \in \mathbf{S}_\infty(M^-, M^+)$ if and only if $\bar{h} \in \mathbf{S}_\infty(-\delta M/2, \delta M/2)$, and $v \in \mathbf{S}_\infty(\epsilon^-, \epsilon^+)$ if and only if $\bar{v} \in \mathbf{S}_\infty(-\delta \epsilon/2, \delta \epsilon/2)$. Similarly, for any $y \in \mathbf{Y}$, let \bar{y} be defined

as in Theorem 3.2, i.e., $\bar{y} := y - UT_n \bar{h} - T_n \bar{v}$. Define further

$$\begin{aligned} \bar{\mathbf{Y}} &:= \{\bar{y} \in \mathbf{R}^n : \bar{y} = UT_n \bar{h} + T_n \bar{v} \text{ for some} \\ &\quad \bar{h} \in \mathbf{S}_\infty(-\delta M/2, \delta M/2), \bar{v} \in \mathbf{S}_\infty(-\delta \epsilon/2, \delta \epsilon/2)\}. \end{aligned}$$

Correspondingly, define the sets

$$\begin{aligned} \bar{\mathbf{P}}(\bar{y}) &:= \{\bar{h} \in \mathbf{S}_\infty(-\delta M/2, \delta M/2) : \bar{y} = UT_n \bar{h} + T_n \bar{v} \\ &\quad \text{for some } \bar{v} \in \mathbf{S}_\infty(-\delta \epsilon/2, \delta \epsilon/2)\} \end{aligned}$$

and $\bar{\mathbf{P}}_n(\bar{y}) := \{T_n \bar{h} : \bar{h} \in \bar{\mathbf{P}}(\bar{y})\}$. It follows that $\bar{y} \in \bar{\mathbf{Y}}$ if and only if $y \in \mathbf{Y}$ and that $T_n \bar{h} \in \bar{\mathbf{P}}_n(\bar{y})$ if and only if $T_n h \in \mathbf{P}_n(y)$. As a result, we have

$$\sup_{y \in \mathbf{Y}} d(\mathbf{P}_n(y), \|\cdot\|) = \sup_{\bar{y} \in \bar{\mathbf{Y}}} d(\bar{\mathbf{P}}_n(\bar{y}), \|\cdot\|).$$

Since $\mathbf{S}_n(-\delta M/2, \delta M/2) \times \mathbf{S}_n(-\delta \epsilon/2, \delta \epsilon/2)$ is a balanced convex set, we may invoke Fact 2.3. This gives

$$\begin{aligned} \sup_{y \in \mathbf{Y}} d(\mathbf{P}_n(y), \|\cdot\|) &= 2 \sup_{T_n \bar{h} \in \bar{\mathbf{P}}_n(\bar{y})} \|T_n \bar{h}\| \\ &= 2 \sup_{T_n \bar{h} \in \bar{\mathbf{P}}_n(T_n \bar{h} + T_n \bar{v})} \|T_n \bar{h} - T_n \bar{h}\| \end{aligned}$$

which establishes (3.10) \square

Theorem 3.3 Let $A_n(y)$ be given by (3.3), and let δM_k and $\delta \epsilon_k$ be defined by (3.4). Then

$$\begin{aligned} \frac{1}{2} \sum_{k=n}^{\infty} \delta M_k + \frac{1}{2} \max_{0 \leq k \leq n-1} \tau_k &\leq \epsilon(A_n; \|\cdot\|_1) \\ &\leq \frac{1}{2} \sum_{k=n}^{\infty} \delta M_k + \sum_{k=0}^{n-1} \tau_k \end{aligned} \quad (3.11)$$

where

$$\tau_k := \min \left\{ \sum_{j=0}^k |a_{k-j}| \delta \epsilon_j, \delta M_k \right\}.$$

Furthermore, A_n will be convergent in ℓ_1 if $\{\delta M_k\}_{k=0}^\infty$ is a monotonically decreasing sequence

Proof From Lemma 3.3, we have

$$\begin{aligned} \frac{1}{2} \sum_{k=n}^{\infty} \delta M_k + \frac{1}{2} \sup_{y \in \mathbf{Y}} d(\mathbf{P}_n(y), \|\cdot\|_1) \\ \leq \epsilon(A_n; \|\cdot\|_1) \leq \frac{1}{2} \sum_{k=n}^{\infty} \delta M_k + \sup_{y \in \mathbf{Y}} d(\mathbf{P}_n(y), \|\cdot\|_1). \end{aligned}$$

It follows from Lemma 3.4 that

$$\begin{aligned} \sup_{y \in \mathbf{Y}} d(\mathbf{P}_n(y), \|\cdot\|_1) &= 2 \max \left\{ \|T_n^{-1} T_n(v - \bar{v})\|_1 : T_n^{-1} T_n(v - \bar{v}) \right. \\ &\quad \left. \in \mathbf{S}_n\left(-\frac{\delta M}{2}, \frac{\delta M}{2}\right), v \in \mathbf{S}_\infty(\epsilon^-, \epsilon^+)\right\} \\ &= 2 \max \left\{ \|T_n^{-1} T_n v\|_1 : \left| \sum_{j=0}^k a_{k-j} v(j) \right| \right. \\ &\quad \left. \leq \frac{\delta M_k}{2}, v \in \mathbf{S}_\infty\left(-\frac{\delta \epsilon}{2}, \frac{\delta \epsilon}{2}\right) \right\} \\ &\leq \sum_{k=0}^{n-1} \min \left\{ \sum_{j=0}^k |a_{k-j}| \delta \epsilon_j, \delta M_k \right\}. \end{aligned}$$

This gives the upper bound in (3.11). The lower bound is established by observing that

$$\begin{aligned} \max \left\{ \|U^{r-1} T_n v\|_1 : \left| \sum_{j=0}^k a_{k-j} v(j) \right| \leq \frac{\delta M_k}{2}, \right. \\ \left. v \in S_\infty \left(-\frac{\delta \epsilon}{2}, \frac{\delta \epsilon}{2} \right) \right\} \\ \geq \frac{1}{2} \max_{0 \leq k \leq n-1} \min \left\{ \sum_{j=0}^k |a_{k-j}| \delta \epsilon_j, \delta M_k \right\}. \end{aligned}$$

To establish the convergence result, define l to be the largest integer $k \in \mathbf{Z}_+$ such that $\delta M_k \geq \delta \epsilon \sum_{j=0}^k |a_j|$. Then, for any $k > l$, $\delta M_k < \delta \epsilon \sum_{j=0}^k |a_j|$. Let $l^* := \min\{l, n\}$. Then, it follows that $l^* \rightarrow \infty$ as $\delta \epsilon \rightarrow 0$ and $n \rightarrow \infty$. Consider now weakening the upper bound in (3.11). We have

$$\begin{aligned} \epsilon(A_n; \|\cdot\|_1) &\leq \frac{1}{2} \sum_{k=n}^{\infty} \delta M_k + \sum_{k=0}^{l^*-1} \min \left\{ \delta \epsilon \sum_{j=0}^k |a_j|, \delta M_k \right\} \\ &< \sum_{k=l^*}^{\infty} \delta M_k + l^* \delta M_{l^*}. \end{aligned}$$

Since $\{\delta M_k\}_{k=0}^{\infty}$ is a monotonically decreasing sequence, it follows from [17, Theorem 3.3.1, p. 61] that $l^* \delta M_{l^*} \rightarrow 0$ as $l^* \rightarrow \infty$. Also, $\sum_{k=l^*}^{\infty} \delta M_k \rightarrow 0$ when $l^* \rightarrow \infty$, because $\{\delta M_k\}_{k=0}^{\infty} \in \ell_1$. Therefore, the last upper bound goes to zero as $\delta \epsilon \rightarrow 0$ and $n \rightarrow \infty$. This completes the proof. \square

Additionally, the upper bound in (3.11) clearly bounds the global identification error $\epsilon(A_n; \|\cdot\|_{H_\infty})$. Hence the convergence condition in Theorem 3.3 is also sufficient for the interpolatory algorithm to be convergent in H_∞ . A similar lower bound will show that the condition $\{\delta M_k\}_{k=0}^{\infty} \in \ell_1$ is necessary for the identification error to be finite in H_∞ . To see this, we first give a condition under which the H_∞ and ℓ_1 norms of a transfer function coincide. The proof for this result is simple and is thus omitted.

Lemma 3.5 Suppose that $h \in \mathbf{X}$ satisfies i) $h(k) \geq 0, \forall k \in \mathbf{Z}_+$, ii) $h(k) \leq 0, \forall k \in \mathbf{Z}_+$, or iii) $h(k)h(k+1) \leq 0, \forall k \in \mathbf{Z}_+$. Then, $\|h(z)\|_{H_\infty} = \|h\|_1$.

The following result gives a lower bound for the identification error in the H_∞ case.

Corollary 3.1 Let $A_n(y)$ be given by (3.3), and let δM_k and $\delta \epsilon_k$ be defined by (3.4). Then

$$\epsilon(A_n; \|\cdot\|_{H_\infty}) \geq \frac{1}{2} \max \left\{ \sum_{k=n}^{\infty} \delta M_k, \max_{0 \leq k \leq n-1} \gamma_k \right\}. \quad (3.12)$$

Proof: Since $\mathbf{P}(y)/\mathbf{P}_n(y)$ is a symmetric set, it follows from [32, p. 11] that

$$\begin{aligned} r(\mathbf{P}(y)/\mathbf{P}_n(y); \|\cdot\|_{H_\infty}) &= \frac{1}{2} d(\mathbf{P}(y)/\mathbf{P}_n(y); \|\cdot\|_{H_\infty}) \\ &\leq \frac{1}{2} d(\mathbf{P}(y)/\mathbf{P}_n(y); \|\cdot\|_1) \\ &= \frac{1}{2} \sum_{k=n}^{\infty} \delta M_k. \end{aligned}$$

However

$$\begin{aligned} d(\mathbf{P}(y)/\mathbf{P}_n(y); \|\cdot\|_{H_\infty}) &= \sup_{h \in \mathbf{P}(y)/\mathbf{P}_n(y)} \|\hat{h}(z) - \hat{p}(z)\|_{H_\infty} \\ &\geq \left\| \sum_{k=n}^{\infty} \delta M_k z^k \right\|_{H_\infty}. \end{aligned}$$

From Lemma 3.5, we have $\|\sum_{k=n}^{\infty} \delta M_k z^k\|_{H_\infty} = \sum_{k=n}^{\infty} \delta M_k$. Therefore, $r(\mathbf{P}(y)/\mathbf{P}_n(y); \|\cdot\|_{H_\infty}) = (1/2) \sum_{k=n}^{\infty} \delta M_k$. This proves the fact that $\epsilon(A_n; \|\cdot\|_{H_\infty}) \geq (1/2) \sum_{k=n}^{\infty} \delta M_k$. Additionally, note that $d(\mathbf{P}(y); \|\cdot\|_{H_\infty}) \geq d(\mathbf{P}(y); \|\cdot\|_2) \geq d(\mathbf{P}_n(y); \|\cdot\|_2)$. It follows analogously as in the proof of Theorem 3.3 that

$$\begin{aligned} \sup_{y \in \mathbf{Y}} d(\mathbf{P}_n(y); \|\cdot\|_2) \\ = 2 \max \left\{ \|U^{r-1} T_n v\|_2 : \left| \sum_{j=0}^k a_{k-j} v(j) \right| \leq \frac{\delta M_k}{2}, \right. \\ \left. v \in S_\infty \left(-\frac{\delta \epsilon}{2}, \frac{\delta \epsilon}{2} \right) \right\} \\ \geq \max_{0 \leq k \leq n-1} \min \left\{ \sum_{j=0}^k |a_{k-j}| \delta \epsilon_j, \delta M_k \right\}. \end{aligned}$$

This establishes that $\epsilon(A_n; \|\cdot\|_{H_\infty}) \geq (1/2) \gamma_k$ for all $k \in \mathbf{Z}_+, n$. The proof is now completed. \square

C. The Central Algorithm and Optimal Error

We now turn to the central algorithm and optimal identification error. Our purpose in this section is to demonstrate, by considering the case of ℓ_1 norm, that the problem of solving the central algorithm and optimal error may, in general, be computationally infeasible, so is the problem of computing exactly the identification error and the diameter of information. This should give certain justification to our pursuit of an interpolatory algorithm and error bounds.

From Lemma 3.2, it is clear that finding the central algorithm in the ℓ_1 case requires computing the center of $\mathbf{P}_n(y)$ and that the optimal identification error corresponds to the radius of the same set. Since $\mathbf{P}_n(y)$ is a convex polytope, it can be expressed in terms of its vertices [35]. Let the vertices of $\mathbf{P}_n(y)$ be denoted by the collection $\{h_1, \dots, h_m\}$, where $h_i \in \mathbf{P}_n(y)$ for $1 \leq i \leq m$. Then

$$\mathbf{P}_n(y) = \left\{ \sum_{i=1}^m \lambda_i h_i : h_i \in \mathbf{P}_n(y), \lambda_i \geq 0, \sum_{i=1}^m \lambda_i = 1 \right\}. \quad (3.13)$$

The following result gives a characterization of $r(\mathbf{P}_n(y); \|\cdot\|_1)$ in terms of the vertices of $\mathbf{P}_n(y)$. The proof is simple and thus is omitted.

Proposition 3.1 Let $\mathbf{P}_n(y)$ be described by (3.13). Then

$$r(\mathbf{P}_n(y); \|\cdot\|_1) = \begin{aligned} &\text{minimize } \delta \\ &\text{subject to } \|h - h_k\|_1 \leq \delta, k=1, 2, \dots, m \end{aligned} \quad (3.14)$$

It is not difficult to see that the minimization problem in (3.14) amounts to a linear program, which, however, requires explicit use of the vertices of $\mathbf{P}_n(y)$. The problem of finding the vertices of a convex polytope has been well studied; see [22], [25], and [34] and the references therein. An important result in [22] states that the number of vertices of a polytope increases exponentially with its dimension. According to this result, the task of generating all the vertices of $\mathbf{P}_n(y)$ and hence the problem of computing the central algorithm and optimal error via the linear program (3.14) becomes computationally

intractable. Alternatively, the complexity in computing the radius $r(\mathbf{P}_n(y), \|\cdot\|_1)$ may also be observed by noting that its computation involves maximizing a convex function over a convex polytope. A recent result [2] in combinatorial theory establishes rigorously that the problem of this class is computationally intractable, in the sense that its computational complexity grows exponentially with the dimension of the polytope. It is clear that the problems of computing exactly the identification error and the diameter of information both belong to this class, and the exact computation of these quantities is also computationally intractable.

IV. IDENTIFICATION WITH IMPULSE AND STEP INPUTS

The general framework developed in the preceding sections will now be examined in the more specific context where the following two important signals will be taken as experimental inputs. The first input is the unit impulse signal, denoted by u_I

$$u_I(k) := \begin{cases} 1 & k = 0 \\ 0 & k \geq 1 \end{cases}$$

The second input is the unit step signal, denoted by u_s

$$u_s(k) := 1, \quad k \in \mathbf{Z}_+$$

The identification problem will first be considered with the general *a priori* information, and then that corresponding to the class of exponentially stable systems.

A. Identification with General *a priori* Information

In using the unit impulse signal as an experimental input, the ℓ_1 identification problem for a class of exponentially stable systems has been studied in detail in [15]. Both the central algorithm and global optimal error were given for the ℓ_1 case in these references. Our development in this section constitutes a generalization to [15], and more importantly, we show that the central algorithm and optimal error are actually the same in both the ℓ_1 and H_∞ cases when the impulse signal is used.

Consider first the set $\mathbf{P}(y)$ corresponding to u_I . Notice trivially that I is an identity matrix. As a result, the set $\mathbf{P}_n(y)$ becomes a hyperrectangle in \mathbf{R}^n

$$\begin{aligned} \mathbf{P}_n(y) &= \{T_n h : \max\{y(k) - \epsilon_k^+, M_k^-\} \leq h(k) \\ &\leq \min\{y(k) - \epsilon_k^-, M_k^+\}, \quad k \in \mathbf{Z}_{+n}\}. \end{aligned}$$

Clearly, the set $\mathbf{P}_n(y)$, and therefore $\mathbf{P}(y)$, are symmetric. For any $k \in \mathbf{Z}_{+n}$, let $h_n(k) = \min\{y(k) - \epsilon_k^-, M_k^+\}$, and $h_I(k) = \max\{y(k) - \epsilon_k^+, M_k^-\}$. Then, the point of symmetry of $\mathbf{P}(y)$ is at $c \in \mathbf{P}(y)$, where $c(k) := (1/2)(h_n(k) + h_I(k))$ for $k \in \mathbf{Z}_{+n}$, and $c(k) := (M_k^+ + M_k^-)/2$ for $k \geq n$. This gives a symmetrically central algorithm.

Theorem 4.1: Let $u = u_I$. Also, let $A_n^*(y)$ be given by

$$[A_n^*(y)]_k := \begin{cases} \frac{1}{2}(h_n(k) + h_I(k)) & k \leq n-1 \\ \frac{M_k^+ + M_k^-}{2} & k \geq n. \end{cases} \quad (4.1)$$

Then, A_n^* is symmetrically central and hence strongly optimal for any norm.

Note that the strong optimality of this algorithm was only established in the ℓ_1 case in [15]. Our derivation shows that it is strongly optimal for any norm. Additionally, our following result gives explicit

expressions of the optimal identification errors in both the ℓ_1 and H_∞ cases.

Theorem 4.2: Let $u = u_I$. Then:

i) for any data record y

$$\begin{aligned} \epsilon^*(y; \|\cdot\|_{H_\infty}) &= \epsilon^*(y; \|\cdot\|_1) \\ &= \frac{1}{2} \sum_{k=n}^{\infty} \delta M_k + \frac{1}{2} \sum_{k=0}^{n-1} (h_n(k) - h_I(k)). \end{aligned} \quad (4.2)$$

ii) Furthermore

$$\begin{aligned} \epsilon^*(\|\cdot\|_{H_\infty}) &= \epsilon^*(\|\cdot\|_1) \\ &= \frac{1}{2} \sum_{k=n}^{\infty} \delta M_k + \frac{1}{2} \sum_{k=0}^{n-1} \min\{\delta \epsilon_k, \delta M_k\}. \end{aligned} \quad (4.3)$$

Proof: We first establish (4.2)-(4.3) for the ℓ_1 case. In this regard, the proof for (4.2) follows from the fact that $A_n^*(y)$ is the point of symmetry of $\mathbf{P}(y)$. It can be readily verified that (4.2) gives the radius of $\mathbf{P}(y)$. Indeed, the radius $r(\mathbf{P}(y), \|\cdot\|_1) = \max_{h \in \mathbf{P}(y)} \|h - A_n^*(y)\|_1$ is achieved at h^* , where $h^*(k) := h_n(k)$ for $k \in \mathbf{Z}_{+n}$ and $h^*(k) := M_k^+$ for $k \geq n$. Since $\mathbf{P}(y)$ is a symmetric set, we have $r(\mathbf{P}(y), \|\cdot\|_1) = (1/2)d(\mathbf{P}(y), \|\cdot\|_1)$ for any y (see, e.g., [32, p. 11]). The proof for (4.3) follows from this observation and using Lemma 3.4. To show that $\epsilon^*(y; \|\cdot\|_{H_\infty}) = \epsilon^*(y; \|\cdot\|_1)$, one simply notes that $h^*(k) - [A_n^*(y)]_k \geq 0$ for all $k \in \mathbf{Z}_+$. The result then follows by applying Lemma 3.5. \square

B. Identification of Exponentially Stable Systems

We now consider a class of exponentially stable systems. The plant *a priori* information corresponding to this class of systems consists of two constants $0 \leq M < \infty$ and $\rho > 1$, so that $-M_k^- = M_k^+ = M/\rho^k$ for all $k \in \mathbf{Z}_+$ [15]. Following [15], we shall also assume a uniform noise bound $\epsilon \geq 0$ for all measurement instants: $-\epsilon_k^- = \epsilon_k^+ = \epsilon$ for all $k \in \mathbf{Z}_+$. As a result, we have $\delta M_k = 2M/\rho^k$, and $\delta \epsilon_k = 2\epsilon$. A direct application of Theorem 4.2 then yields the following simplified expression of the optimal global error in the impulse input case. The result for ℓ_1 case is precisely the one given in [15].

Corollary 4.1 Let $u = u_I$. Denote $n^* = (1/\ln \rho) \ln(M/\epsilon)$, and $n_I = \min\{n, \lceil n^* \rceil\}$. Then

$$\epsilon^*(\|\cdot\|_{H_\infty}) = \epsilon^*(\|\cdot\|_1) = n_I \epsilon + \frac{M}{\rho^{n_I-1}(\rho-1)}. \quad (4.4)$$

In the step input case, we give an explicit expression of the global diameter of information for both the ℓ_1 and H_∞ identification problems, hence providing an estimate tight to within a factor of two of the global optimal error. We summarize this result in the following theorem.

Theorem 4.3. Let $u = u_s$. Let also

$$n_s := \begin{cases} 0 & M \leq \epsilon \\ \min\{n, \lceil n^* \rceil\} & \text{otherwise} \end{cases} \quad (4.5)$$

where $n^* := (1/\ln \rho) \ln(M/2\epsilon)$. Then

$$\begin{aligned} \sup_{y \in \mathbf{Y}} d(\mathbf{P}(y), \|\cdot\|_{H_\infty}) &= \sup_{y \in \mathbf{Y}} d(\mathbf{P}(y), \|\cdot\|_1) \\ &= 2 \begin{cases} n_s \epsilon + \frac{M}{\rho^{n_s-1}(\rho-1)} & n_s \leq 1 \\ (2n_s - 1)\epsilon + \frac{M}{\rho^{n_s-1}(\rho-1)} & n_s > 1. \end{cases} \end{aligned} \quad (4.6)$$

Proof: Again, we first prove the result for the ℓ_1 case. In this regard, we notice that for $u = u_s$,

$$U^{n-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -1 & 1 \end{bmatrix}.$$

Therefore, $h \in \mathbf{P}(0)$ if and only if $|h(k)| \leq M/\rho^k$ for all $k \in \mathbf{Z}_+$, and $h(0) = v(0)$, $h(k) = v(k) - v(k-1)$ for $k \in \mathbf{Z}_{+n}$, where $|v(k)| \leq \epsilon$. It follows from Fact 2.3 that

$$\begin{aligned} & \frac{1}{2} \sup_{y \in Y} d(\mathbf{P}(y), \|\cdot\|_1) \\ &= \max \left\{ \sum_{k=0}^{n-1} |v(k) - v(k-1)| : |v(k)| \leq \epsilon, \right. \\ & \quad \left. |v(k) - v(k-1)| \leq \frac{M}{\rho^k} \right\} + \sum_{k=n}^{\infty} \frac{M}{\rho^k} \\ &\leq \min\{\epsilon, M\} + \sum_{k=1}^{n-1} \min\left\{2\epsilon, \frac{M}{\rho^k}\right\} + \sum_{k=n}^{\infty} \frac{M}{\rho^k}. \end{aligned}$$

If $n_s = 0$ (i.e., $M \leq \epsilon$) or $n_s = 1$ ($M > \epsilon$, $2\epsilon \leq M/\rho$), then we have

$$\begin{aligned} & \frac{1}{2} \sup_{y \in Y} d(\mathbf{P}(y), \|\cdot\|_1) \leq n_s \epsilon + \sum_{k=n_s}^{\infty} \frac{M}{\rho^k} \\ &= n_s \epsilon + \frac{M}{\rho^{n_s-1}(\rho-1)}. \end{aligned}$$

It is easy to see that the equality can actually be attained by the sequence $\{v(k)\}_{k=0}^{n-1}$

$$v(k) := \begin{cases} v(0) & k=0 \\ v(k-1) + (-1)^k M/\rho^k & 1 \leq k \leq n-1 \end{cases} \quad (4.7)$$

where $v(0) = M$ for $n_s = 0$, and $v(0) = \epsilon$ for $n_s = 1$. Hence (4.6) holds for the case $n_s = 0$ and $n_s = 1$. Consider now the case $n_s > 1$. It follows that

$$\begin{aligned} & \frac{1}{2} \sup_{y \in Y} d(\mathbf{P}(y), \|\cdot\|_1) \leq \epsilon + 2(n_s - 1)\epsilon + \sum_{k=n_s}^{\infty} \frac{M}{\rho^k} \\ &= (2n_s - 1)\epsilon + \frac{M}{\rho^{n_s-1}(\rho-1)}. \end{aligned}$$

Again, equality can be attained by the sequence $\{v(k)\}_{k=0}^{n-1}$, constructed as

$$v(k) := \begin{cases} (-1)^k \epsilon & k \leq n_s \\ v(k-1) + (-1)^k M/\rho^k & n_s \leq k \leq n-1 \end{cases} \quad (4.8)$$

Hence (4.6) also holds for the case $n_s > 1$. This completes the proof for the ℓ_1 case. To show that the global diameters are identical in both the ℓ_1 and H_∞ cases, we note that the sequences constructed in (4.7) and (4.8) lead to $h(k)h(k+1) \leq 0$, where $h(k) = v(k) - v(k-1)$ for all $k \in \mathbf{Z}_{+n}$. The result then follows from Lemma 3.5. \square

V. CONCLUSION

In this paper we have formulated and solved a worst-case/deterministic system identification problem. Our main effort has been concentrated on developing an interpolatory algorithm and bounds for the identification error. An important reason in our pursuit of the interpolatory algorithm and error bounds has been that the optimal algorithm and error is generally very difficult to compute. In contrast, the interpolatory algorithm, albeit suboptimal, can be computed easily by solving a linear program.

The proposed algorithm is "tuned" [12] to the plant and noise *a priori* information. It is possible, as in [15], to construct an algorithm tuned to only the plant *a priori* information. Note that in the present formulation the past input has been assumed to be zero. Nevertheless, this assumption can be relaxed, and the present results can be extended to the more general setting, as evidenced by an extension in [11], which appears to be directly built on the present approach and techniques. Future extensions building on this work may also be pursued by considering mixed parametric and nonparametric identification problems [19], [18], [9]. Finally, we note the recent work in this area on input design and sample complexity in e.g., [5], [29], [16], and [1].

ACKNOWLEDGMENT

The authors would like to thank Profs. M. Milanese and J. P. Norton for calling [25] and [34] to their attention, respectively, and Profs. J. R. Partington and G. Gu for pointing out an error in an earlier draft of this paper.

REFERENCES

- [1] G. Belforte and T. T. Tay, "Optimal input design for worst-case system identification in $\ell_1/\ell_2/\ell_\infty$," *Syst. Contr. Lett.*, vol. 20, pp. 273-278, 1993.
- [2] H. L. Bodlaender, P. Gritzmann, V. Klee, and J. Van Leeuwen, "Computational complexity of norm maximization," *Combinatorica*, vol. 10, no. 2, pp. 203-225, 1990.
- [3] J. Chen, C. N. Nett, and M. K. H. Fan, "Optimal nonparametric system identification from arbitrary corrupt finite time series: a control-oriented approach," in *Proc. 1992 Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 279-285.
- [4] —, "Worst case identification in H_∞ : Validation of *a priori* information, essentially optimal algorithms, and error bounds," in *Proc. 1992 Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 251-257.
- [5] M. A. Dahleh, T. Theodopoulos, and J. N. Tsitsiklis, "The sample complexity of worst-case identification of FIR linear systems," *Syst. Contr. Lett.*, vol. 20, no. 3, Mar. 1993.
- [6] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*. New York: Academic, 1975.
- [7] P. Dorato, Ed., *Robust Control*. New York: IEEE Press, 1987.
- [8] P. Dorato and R. Yedavalli, Eds., *Recent Advances in Robust Control*. New York: IEEE Press, 1990.
- [9] N. Elia and M. Milanese, "Worst-case ℓ_1 identification using mixed parametric-nonparametric models," in *Proc. 32nd IEEE Conf. Dec. Contr.*, San Antonio, TX, Dec. 1993, pp. 545-550.
- [10] G. Gu and P. P. Khargonekar, "A class of algorithms for identification in H_∞ ," *Automatica*, vol. 28, no. 3, pp. 229-312, Mar. 1992.
- [11] R. G. Hakvoort, "Worst-case system identification in ℓ_1 : error bounds, optimal models, and model reduction," in *Proc. 31st IEEE Conf. Dec. Contr.*, Tucson, AZ, Dec. 1992, pp. 499-504.
- [12] A. J. Helmicki, C. A. Jacobson, and C. N. Nett, "Control oriented system identification: a worst-case/deterministic approach in H_∞ ," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1163-1176, Oct. 1991.
- [13] R. B. Holmes, *A Course on Optimization and Best Approximation* (Lecture Notes in Mathematics), vol. 257. Berlin: Springer-Verlag, 1972.
- [14] I. S. Iohvidov, *Hankel and Toeplitz Matrices and Forms*. Boston: Birkhäuser, 1982.
- [15] C. A. Jacobson, C. N. Nett, and J. R. Partington, "Worst case system identification in ℓ_1 : optimal algorithms and error bounds," *Syst. Contr. Lett.*, vol. 19, pp. 419-424, 1992.
- [16] B. Kacwicz and M. Milanese, "Optimal finite-sample experiment design in the worst-case ℓ_1 system identification," in *Proc. 31st IEEE Conf. Dec. Contr.*, Tucson, AZ, Dec. 1992, pp. 56-61.
- [17] K. Knopp, *Infinite Sequences and Series*. New York: Dover, 1956.
- [18] R. L. Kosut, M. Lau, and S. Boyd, "Identification of systems with parametric and nonparametric uncertainty," in *Proc. 1990 Amer. Contr. Conf.*, San Diego, CA, May 1990, pp. 2412-2417.
- [19] J. M. Krause and P. P. Khargonekar, "Parameter identification in the presence of nonparametric uncertainty," *Automatica*, vol. 26, no. 1, pp. 113-123, Jan. 1990.

- [20] P. M. Mäkilä, "Robust identification and Galois sequences," *Int. J. Contr.*, vol. 54, no. 5, pp. 1189–1200, 1991.
- [21] P. M. Mäkilä and J. R. Partington, "Robust approximation and identification in H_∞ ," in *Proc. 1991 Amer. Contr. Conf.*, Boston, MA, June 1991, pp. 70–76.
- [22] T. H. Maheiss and D. S. Rubin, "A survey and comparison of methods for finding all vertices of convex polyhedral sets," *Math. Oper. Research*, vol. 5, no. 2, pp. 167–185, 1980.
- [23] M. Milanese and R. Tempo, "Optimal algorithms theory for robust estimation and prediction," *IEEE Trans. Automat. Contr.*, vol. AC-30, no. 8, pp. 730–738, Aug. 1985.
- [24] M. Milanese, R. Tempo, and A. Vicino, "Strongly optimal algorithms and optimal information in estimation problems," *J. Complexity*, vol. 2, pp. 78–94, 1986.
- [25] M. Milanese and A. Vicino, "Optimal estimation theory for dynamic systems with set membership uncertainty: An overview," *Automatica*, vol. 27, no. 6, pp. 997–1009, 1991.
- [26] K. G. Murty, *Linear Programming*. New York: Wiley, 1983.
- [27] J. R. Partington, "Robust identification and interpolation in H_∞ ," *Int. J. Contr.*, vol. 54, pp. 1281–1290, 1991.
- [28] K. Poolla, P. P. Khargonekar, A. Tikku, J. Krause, and K. M. Nagpal, "A time domain approach to model validation," in *Proc. 1992 Amer. Contr. Conf.*, Chicago, IL, June 1992, pp. 313–317.
- [29] K. Poolla and A. Tikku, "On the time complexity of worst-case system identification," in *Proc. 1993 Amer. Contr. Conf.*, San Francisco, CA, June 1993.
- [30] R. Smith and J. C. Doyle, "Model validation: A connection between robust control and identification," *IEEE Trans. Automat. Contr.*, vol. 37, no. 7, pp. 942–952, July 1992.
- [31] J. F. Traub, G. W. Wasilkowski, and H. Wozniakowski, *Information-Based Complexity*. New York: Academic, 1988.
- [32] J. F. Traub and H. Wozniakowski, *A General Theory of Optimal Algorithms*. New York: Academic, 1980.
- [33] D. C. N. Tse, M. A. Dahleh, and J. N. Tsitsiklis, "Optimal asymptotic identification under bounded disturbances," *IEEE Trans. Automat. Contr.*, vol. 38, no. 8, pp. 1176–1190, Aug. 1993.
- [34] E. Walter and H. Piet-Lahanier, "Estimation of parameter bounds from bounded-error data: A survey," *Math. and Computers in Simulation*, vol. 32, pp. 449–468, 1990.
- [35] V. A. Yemelichev, M. M. Kovalev, and M. K. Kravtsov, *Polytopes, Graphs and Optimization*. Cambridge, UK: Cambridge Univ. Press, 1984.
- [36] "Special Issue on System Identification for Robust Control Design," *IEEE Trans. Automat. Contr.*, vol. 37, no. 7, July 1992.

Input Saturation and Global Stabilization of Nonlinear Systems via State and Output Feedback

Wei Lin

Abstract—For multi-input multi-output nonlinear systems whose free dynamics are Lyapunov stable, we show how the problem of global stabilization via dynamic output feedback can be solved by using the technique of input saturation. The power of this technique is also illustrated by solving the problem of global stabilization via bounded state feedback for affine nonlinear systems with stable unforced dynamics. Analogous results are established for discrete-time nonlinear systems.

Manuscript received February 1, 1994; revised July 28, 1994. This work was supported in part by grants from AFOSR and NSF.

The author is with the Department of Systems Science and Mathematics, Washington University, St. Louis, MO 63130 USA.

IEEE Log Number 9408791.

I. INTRODUCTION

Consider multi-input multi-output (MIMO) nonlinear systems described by differential equations of the form

$$\dot{x} = Ax + \sum_{i=1}^m g_i(x)u_i \quad (1.1)$$

$$y = Cx \quad (1.2)$$

where $x \in \mathbb{R}^n$ is the state, $u_i \in \mathbb{R}$, $1 \leq i \leq m$, are the control inputs and $y \in \mathbb{R}^s$ is the system output, $A \in \mathbb{R}^{n \times n}$ and $C \in \mathbb{R}^{s \times n}$ are constant matrices, and $g_i: \mathbb{R}^n \rightarrow \mathbb{R}^n$, $1 \leq i \leq m$, are smooth functions. In this paper, we assume that the nonlinear systems under consideration, which include bilinear systems as a particular case, satisfy the following hypothesis:

H1) There exists a positive definite matrix P such that

$$A^T P + P A \leq 0.$$

In other words, the unforced dynamic system of (1.1) is assumed to be Lyapunov stable.

The purpose of this paper is to study the following two problems of global stabilization for a nonlinear system (1.1)–(1.2) with Hypothesis H1).

Problem 1: When is there a smooth state feedback or a bounded state feedback control law which renders the equilibrium $x = 0$ of system (1.1) globally asymptotically stable?

In the case where the state x is not available for measurement but only the output y is measurable, the following problem becomes more realistic and rather important.

Problem 2: When does there exist a dynamic compensator

$$\begin{aligned} \dot{\xi} &= \eta(\xi, y), \quad \eta(0, 0) = 0 \\ u &= \theta(\xi), \quad \theta(0) = 0 \end{aligned} \quad (1.3)$$

such that the closed-loop system (1.1)–(1.3) is globally asymptotically stable at the equilibrium $(x, \xi) = (0, 0)$?

In the case of single-input single-output (SISO) bilinear systems (i.e., $g(x) = Bx + N$), these two problems have been investigated by Gauthier–Kupka in [7], which improves the nonlinear separation principle initially developed by Vidyasagar [17] in the bilinear case. More specifically, Gauthier and Kupka [7] develop sufficient conditions for global stabilization of SISO bilinear systems via smooth state feedback by using the Jurdjevic–Quinn technique [8]. Based on the Jurdjevic–Quinn type feedback control law, they then present a "small stabilizing control law." This, in turn, leads to a separation principle for a class of SISO bilinear systems whose free dynamics are dissipative. The result of [7] is nice, and the proof is clever but somewhat unsatisfactory because it is only applicable to a SISO system. In addition, it requires that for $W \triangleq \{x: x^T(A^T P + P A)x = 0\} \cap \{x: x^T P(Bx + N) = 0\}$

$$\begin{aligned} \dim \operatorname{span} \{Ax, \operatorname{ad}_{Ax}^0(Bx + N), \dots, \operatorname{ad}_{Ax}^k(Bx + N), \dots\} \\ = n \quad \forall x \in W - \{0\} \end{aligned} \quad (1.4)$$

i.e., the ad-condition is satisfied on $W - \{0\}$.

Note that requirement (1.4) is too strong even for a linear system. More explicitly, it is easy to see that for the linear system [12]

$$\Sigma_L: \dot{x} = Ax + bu = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 2 \\ 0 & 0 & -1 \end{bmatrix} x + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u$$

$$\dim \text{span}\{Ax, b, Ab, A^2b\}$$

$$= \dim \text{span}\left\{ \begin{bmatrix} 0 \\ -x_2 + 2x_3 \\ -x_3 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right\} \leq 2.$$

Thus, ad-condition (1.4) can never be satisfied. If, however, we use Corollary A.2, which is a slight generalization of Theorem 1 given in [12], a straightforward calculation shows that for $V(x) = x_1^2 + x_2^2 + x_3^2$

$$\Omega_{\Sigma} = \{x \in \mathbb{R}^3 : x_2 = x_3 = 0\}$$

and

$$S_{\Sigma} = \{x \in \mathbb{R}^3 : x_1 = 0\}.$$

Hence, $\Omega_{\Sigma} \cap S_{\Sigma} = \{0\}$. Thus Σ_L is stabilizable by $u = -L_b V(x) = -2x_1$.

In this paper, we address Problems 1 and 2 for MIMO nonlinear control systems (1.1)–(1.2), with the property H1). In particular, a solution to these problems is proposed. The approach employed in the paper is based on Lyapunov analysis, LaSalle's invariance principle, and the center manifold theory [5], in the spirit of [7]. Our solution to Problem 2 depends heavily on the technique of input saturation, which has recently obtained a strong renewed interest in the linear control literature [13], [15]–[16]. The power of this technique is first illustrated by solving the problem of global stabilization via bounded state feedback for affine nonlinear control systems. The result incorporates and generalizes a number of well-known stabilization theorems [1], [6], [8]–[9], [12], and [14] proposed in the nonlinear control literature for global asymptotic stabilization of certain classes of affine systems. Since the material is somehow independent of the theme of this paper, we put it in Appendix A. Appendix B includes a useful boundedness criterion for perturbed systems. Section II deals with Problems 1 and 2 for MIMO continuous-time nonlinear systems of the form (1.1)–(1.2) and we show how the ad-condition can be removed. Discrete-time counterparts of systems (1.1)–(1.2) are treated in Section III. Section IV includes the concluding remarks of this paper.

The preliminary version of this work was presented at the 33rd IEEE Conference on Decision and Control [10].

II. SOLUTIONS OF GLOBAL STABILIZATION PROBLEMS

In this section, we present a solution to Problems 1 and 2 stated in the previous section. To describe our main results in this paper, we need to recall some notations related to nonlinear systems of the form (1.1).

For smooth vector fields Ax and $g_i(x)$, $1 \leq i \leq m$, on \mathbb{R}^n , $[Ax, g_i(x)]$ denotes their Lie bracket and we define inductively

$$\begin{aligned} ad_{Ax}^0 g_i(x) &= g_i(x), \dots, ad_{Ax}^{k+1} g_i(x) \\ &= [Ax, ad_{Ax}^k g_i(x)], \quad k = 0, 1, \dots \end{aligned}$$

for $i = 1, \dots, m$. If $V: \mathbb{R}^n \rightarrow \mathbb{R}$ is a C^r ($r \geq 1$) smooth function, the Lie derivative of V with respect to the vector field Ax is defined by $L_{Ax} V(x) = \frac{\partial V}{\partial x} Ax$. In a similar manner, higher order Lie derivatives can be defined inductively as follows

$$\begin{aligned} L_{Ax}^0 V(x) &= V(x), \dots, L_{Ax}^{k+1} V(x) \\ &= L_{Ax} (L_{Ax}^k V(x)) \quad 1 \leq k \leq r-1. \end{aligned}$$

With the vector fields $Ax, g_1(x), \dots, g_m(x)$ we associate the distribution

$$D = \text{span}\{ad_{Ax}^k g_i(x) : 0 \leq k \leq n-1, 1 \leq i \leq m\}. \quad (2.1)$$

Let Ω and S denote the sets associated with $V(x) \triangleq \frac{1}{2} x^T P x$, $P > 0$

$$\left. \begin{aligned} \Omega &\triangleq \{x \in \mathbb{R}^n : L_{Ax}^k V(x) = 0 = x^T P A^k x, \\ &\quad 1 \leq k \leq r\} \\ S &\triangleq \{x \in \mathbb{R}^n : L_{Ax}^k L_{\tau}(x^T P x) = 0 \\ &\quad \forall \tau \in D, 0 \leq k \leq r-1\} \end{aligned} \right\}. \quad (2.2)$$

With these notations in mind, we can easily deduce from Appendix A sufficient conditions for the solution of global stabilization via smooth state feedback or bounded state feedback for continuous-time multi-input nonlinear systems of form (1.1).

Theorem 1: Suppose a continuous-time multi-input nonlinear system (1.1) satisfies Hypothesis H1). If $\Omega \cap S = \{0\}$, then

- i) the equilibrium $x = 0$ of system (1.1) is globally asymptotically stabilized by the bounded state feedback control law

$$\begin{aligned} u_i^*(x) &= -\text{sgn}(x^T P g_i(x)) \\ &\quad \times \min(|x^T P g_i(x)|, \alpha), \quad 1 \leq i \leq m \end{aligned} \quad (2.3)$$

where α is an arbitrarily positive real number.

As an immediate consequence of (2.3),

- ii) system (1.1) is globally asymptotically stabilizable via the smooth state feedback control law

$$u = -[g_1(x), \dots, g_m(x)]^T P x \triangleq -g^T(x) P x. \quad (2.4)$$

Proof: Let $V(x) = \frac{1}{2} x^T P x$ with $P > 0$. Conclusion i) follows from Theorem A.1 proposed in the appendix; ii) is a consequence of Corollary A.2.

In the case where the state x of system (1.1) is not available for measurement, the feedback control schemes proposed in Theorem 1 cannot be applied directly, thus it is necessary and important to develop sufficient conditions for Problem 2 to be solvable. In the remainder of this section, we assume that the pair (A, C') of the input-output nonlinear system (1.1)–(2.2) is observable. Then, it can be shown that a dynamic compensator of the form (1.3), which consists of a "Luenberger-type" nonlinear observer and a bounded state feedback control strategy, solves Problem 2 for MIMO nonlinear systems (1.1)–(1.2). \square

Theorem 2: Consider continuous-time MIMO nonlinear systems of the form (1.1)–(1.2) satisfying Hypothesis H1). Suppose the pair (A, C') is observable and $\Omega \cap S = \{0\}$. Assume that the functions $g_i(x)$, $1 \leq i \leq m$, are global Lipschitz $\forall x \in \mathbb{R}^n$. Then for any sufficiently small $\alpha > 0$, a dynamic compensator

$$\left. \begin{aligned} \dot{\xi} &= A\xi + \sum_{i=1}^m g_i(\xi) u_i + K(y - C\xi) \\ u_i &= u_i^*(\xi), \quad 1 \leq i \leq m \end{aligned} \right\} \quad (2.5)$$

is such that the closed-loop system formed by (1.1), (1.2), and (2.5) is globally asymptotically stable at the equilibrium $(x, \xi) = (0, 0)$, where $u_i^*(\cdot)$ is given by (2.3) and the gain matrix K is designed in such a way that $A - KC$ is Hurwitz.

Proof: Let e denote the error signal

$$e(t) = \xi(t) - x(t).$$

Then, the closed-loop system can be represented as

$$\dot{e} = (A - KC)e + \sum_{i=1}^m (g_i(\xi) - g_i(x))u_i^\alpha(\xi) \quad (2.6)$$

$$\dot{\xi} = A\xi + \sum_{i=1}^m g_i(\xi)u_i^\alpha(\xi) - KCe. \quad (2.7)$$

Since $A - KC$ is Hurwitz, there exists a unique positive definite matrix R satisfying the Lyapunov equation

$$(A - KC)^T R + R(A - KC) = -I.$$

Thus a straightforward calculation proves that

$$\begin{aligned} \frac{d}{dt}(e^T Re) &= -e^T e + 2e^T R \sum_{i=1}^m (g_i(\xi) - g_i(x))u_i^\alpha(\xi) \\ &\leq -e^T e \left(1 - 2\alpha \|R\| \sum_{i=1}^m \beta_i\right) \end{aligned}$$

because $|u_i^\alpha(\xi)| \leq \alpha$ and $\|g_i(x) - g_i(\xi)\| \leq \beta_i \|x - \xi\|$ for $\alpha > 0$ and $\beta_i > 0$, $i = 1, \dots, m$. Choosing $\alpha \leq \frac{1}{4\|R\| \sum_{i=1}^m \beta_i}$, yields

$$\frac{d}{dt}(e^T Re) \leq -\frac{1}{2}e^T e \leq -\frac{1}{2\lambda_{\min}(R)}e^T Re.$$

This, in turn, implies

$$\begin{aligned} \lambda_{\min}(R)\|e\|^2 &\leq e^T Re \\ &\leq \exp\left(-\frac{t}{2\lambda_{\min}(R)}\right)\exp(e(0)^T Re(0)). \end{aligned}$$

Hence

$$\|e(t)\| \leq M \exp(-\beta t) \quad \text{for } M > 0 \text{ and } \beta > 0. \quad (2.8)$$

Let $d(t) = -KCe(t)$ in (2.7). Recall that by Theorem 1 system (2.7) with $d(t) = 0$ is globally asymptotically stable at $\xi = 0$. In particular, for $V(\xi) \triangleq \xi^T P\xi$ and $f(\xi) \triangleq A\xi + \sum_{i=1}^m g_i(\xi)u_i^\alpha(\xi)$

$$L_f V(\xi) = f^T(\xi)P\xi + \xi^T P f(\xi) \leq 0.$$

Moreover, it is easy to check that in this case

$$V(\xi) = \xi^T P\xi \geq \lambda_{\min}(P)\|\xi\|^2, \quad \left\|\frac{\partial V}{\partial \xi}\right\| \leq 2\|P\|\|\xi\|$$

and

$$\int_0^\infty \|d(t)\| dt \leq \int_0^\infty M\|KC\|\exp(-\beta t) dt < +\infty.$$

With this in mind, we apply Theorem B.1 given in Appendix B to system (2.7) and then conclude that $\|\xi(t)\| \leq N \forall t \geq 0$, for some $N > 0$.

So far, we have shown that every trajectory $(e(t), \xi(t))$ of the closed-loop system (2.6)–(2.7) is bounded $\forall t \geq 0$ and $(e(0), \xi(0)) \in \mathbb{R}^n \times \mathbb{R}^n$. To show $(e, \xi) = (0, 0)$ is a global asymptotically stable equilibrium of (2.6), we let $(e(t), \xi(t))$ be a trajectory of system (2.6)–(2.7) with the initial value $(e(0), \xi(0))$. Let r^0 denote its ω -limit set. Clearly, r^0 is nonempty, compact, and invariant because $(e(t), \xi(t))$ is bounded $\forall t \geq 0$. In addition, it follows from (2.8) that $\lim_{t \rightarrow \infty} e(t) = 0$. Therefore, any point in r^0 must be a pair of the form $(0, \bar{\xi}(t))$. Let $(0, \bar{\xi}) \in r^0$ and

$(0, \bar{\xi}(t))$ be the corresponding trajectory. Obviously, this trajectory is characterized by the following differential equation

$$\dot{\bar{\xi}}(t) = A\bar{\xi}(t) + \sum_{i=1}^m g_i(\bar{\xi}(t))u_i^\alpha(\bar{\xi}(t)) \quad (2.9)$$

which has been proved to be globally asymptotically stable at $\xi = 0$. In other words, the global asymptotic behavior of the closed-loop system (2.6)–(2.7) at $(e, \xi) = (0, 0)$ is completely determined by the flow on the invariant manifold governed by system (2.9) [5]. Since the latter is globally asymptotically stable, so is the closed-loop system (2.6)–(2.7).

As an immediate consequence, we have the following global separation principle for MIMO bilinear systems.

Corollary 1: Consider a continuous-time MIMO bilinear system

$$\begin{cases} \dot{x} = Ax + \sum_{i=1}^m (B_i x + N_i) u_i \\ y = Cx \end{cases} \quad (2.10)$$

Suppose Assumption H1) holds. Suppose the pair (A, C) is observable and $\Omega \cap S = \{0\}$. Then a dynamic compensator of the form (2.5) with $g_i(\xi) = B_i \xi + N_i$, $1 \leq i \leq m$, renders the closed-loop system (2.5)–(2.10) globally asymptotically stable.

III. THE DISCRETE-TIME CASE

The goal of this section is to show how discrete counterparts of the stabilization results established in Section II can be carried out for a class of MIMO discrete-time nonlinear systems of the form

$$x(k+1) = Ax(k) + \sum_{i=1}^m g_i(x(k))u_i(k) \quad (3.1)$$

$$y(k) = Cx(k) \quad (3.2)$$

whose free dynamic system $x(k+1) = Ax(k)$ is Lyapunov stable, so that there is a positive definite matrix P satisfying

$$H2) \quad A^T P A - P \leq 0.$$

Unlike the case of linear systems, the generalization from continuous time to discrete time in the nonlinear context is infamous for its bifurcation into two cases—the routine extension and the technically impossible with relatively few exceptions. This is primarily caused by a nonlinear nature. In [3], we have pointed out that one cannot transport the stabilization results via smooth state feedback for continuous-time affine systems ([1], [6], [8]–[9], [12], [14]) to the discrete-time case. The reason why this is impossible has been explained in [2]–[4] from a passive system point of view. In what follows, on the basis of our previous work [3], we construct a bounded state feedback control law by the use of a modified saturation design technique. This, in turn, leads to a global separation principle for discrete-time nonlinear systems (3.1)–(3.2).

To begin with, we introduce the sets Ω_d and S_d which correspond to the sets defined in (2.2)

$$\begin{cases} \Omega_d = \{x \in \mathbb{R}^n \mid (A^t x)^T (A^T P A - P) A^t x = 0, \\ \quad t = 0, 1, \dots\} \\ S_d = \{x \in \mathbb{R}^n \mid (A^{t+1} x)^T P g_i(A^t x) = 0, \\ \quad 1 \leq i \leq m, t = 0, 1, \dots\} \end{cases} \quad (3.3)$$

In addition, set

$$g(x) = [g_1(x), \dots, g_m(x)]. \quad (3.4)$$

We are now in the position to state and prove our main result in this section. To keep the exposition simple and highlight the central idea behind our design approach, we first study the single-input case.

Theorem 3: Consider a single-input discrete-time nonlinear system (3.1) with the property H2). Suppose $\Omega_d \cap S_d = \{0\}$. Then

- i) the equilibrium $x = 0$ of system (3.1) is globally asymptotically stabilized by the bounded state feedback control law

$$u_\alpha(x) = -\frac{\min(|g^T(x)PAx|, \alpha) \operatorname{sgn}(g^T(x)PAx)}{1 + \frac{1}{2}g^T(x)Pg(x)}, \quad (3.5)$$

for any $\alpha > 0$.

- ii) As a consequence of i), the equilibrium $x = 0$ of (3.1) is globally asymptotically stabilized by the smooth state feedback control law

$$u(x) = -\frac{g^T(x)PAx}{1 + \frac{1}{2}g^T(x)Pg(x)}. \quad (3.6)$$

Proof: Consider a proper Lyapunov function $V(x) = \frac{1}{2}x^T Px$. Then the difference of $V(x)$ along the trajectory of the closed-loop system (3.1)–(3.5) is

$$\begin{aligned} \Delta V(x(k)) &= V(x(k+1)) - V(x(k)) \\ &= \frac{1}{2}x^T(k) \left((A^T P A - P) \right) x(k) \\ &\quad + u_\alpha^T(k) g^T(x(k)) P A x(k) \\ &\quad + \frac{1}{2} u_\alpha^T(k) g^T(x(k)) P g(x(k)) u_\alpha(k). \end{aligned} \quad (3.7)$$

By Assumption H2), it is clear that

$$\begin{aligned} \Delta V(x) &\leq u_\alpha^T g^T(x) P A x \\ &\quad + u_\alpha^T \left(1 + \frac{1}{2} g^T(x) P g(x) \right) u_\alpha - u_\alpha^T u_\alpha. \end{aligned} \quad (3.8)$$

Substituting the feedback control law (3.5) into (3.8) yields

$$\begin{aligned} \Delta V(x) &\leq -u_\alpha^2 - \frac{\min(|g^T(x)PAx|, \alpha) \operatorname{sgn}(g^T(x)PAx)}{1 + \frac{1}{2}g^T(x)Pg(x)} \\ &\quad \times g^T(x)PAx + \frac{[\min(|g^T(x)PAx|, \alpha)]^2}{1 + \frac{1}{2}g^T(x)Pg(x)}. \end{aligned}$$

Note that $|x| = x \operatorname{sgn} x$. With this in mind, we deduce from the inequality above that

$$\begin{aligned} \Delta V(x) &\leq -u_\alpha^2 - \frac{\min(|g^T(x)PAx|, \alpha)}{1 + \frac{1}{2}g^T(x)Pg(x)} \\ &\quad \times \left(|g^T(x)PAx| - \min(|g^T(x)PAx|, \alpha) \right) \leq 0. \end{aligned} \quad (3.9)$$

This proves that the closed-loop system (3.1)–(3.5) is globally stable at $x = 0$. To show global asymptotic stability, we need to distinguish two cases. First, suppose $\min(|g^T(x)PAx|, \alpha) = \alpha$. Then it follows from (3.9) that

$$\Delta V(x) \leq -u_\alpha^2 = -\frac{\alpha^2}{(1 + \frac{1}{2}g^T(x)Pg(x))^2} < 0$$

which implies that the origin of the closed-loop system is globally asymptotically stable. In the case of $\min(|g^T(x)PAx|, \alpha) = |g^T(x)PAx|$, by LaSalle's theorem, all trajectories of the closed-loop system approach to the largest invariant set I contained in the zero locus of $\Delta V(x)$. From (3.7) and (3.9), we see that $\Delta V(x) = 0$ implies for all $k = 0, 1, \dots$

$$u_\alpha(k) = 0, \quad x^T(k) (A^T P A - P) x(k) = 0$$

and

$$g^T(x(k)) P A x(k) = 0.$$

This, in turn, results in $\forall x(0) = x$

$$(A^k x)^T (A^T P A - P) A^k x = 0 \quad \text{and} \quad g^T(A^k x) P A^{k+1} x = 0.$$

In view of the assumption $\Omega \cap S = \{0\}$, we conclude that $I = \{x \in \mathbb{R}^n : \Delta V(x) = 0\} = \{0\}$. This completes the proof of Theorem 3 i). Let $\alpha = +\infty$, conclusion ii) follows from i) immediately.

Similar to the continuous-time case, we can establish a global separation principle for a SISO discrete-time nonlinear system of the form (3.1)–(3.2), by using the bounded state feedback control strategy proposed in Theorem 3.

Theorem 4: Consider a discrete-time SISO nonlinear system (3.1)–(3.2) which satisfies Hypothesis H2). Suppose the pair (A, C) is observable and $\Omega_d \cap S_d = \{0\}$. Suppose the function $g(x)$ is global Lipschitz on \mathbb{R}^n . Then for any sufficiently small $\alpha > 0$, a Luenberger-observer-like based output feedback control law

$$\left. \begin{aligned} \xi(k+1) &= A\xi(k) + g(\xi(k))u_\alpha(k) + K(g(k) - C\xi(k)) \\ u_\alpha(k) &= -\frac{\min(|g^T(\xi(k))PA\xi(k)|, \alpha)}{1 + \frac{1}{2}g^T(\xi(k))Pg(\xi(k))} \\ &\quad \times \operatorname{sgn}(g^T(\xi(k))PA\xi(k)) \end{aligned} \right\} \quad (3.10)$$

renders the equilibrium $(x, \xi) = (0, 0)$ of the closed-loop system (3.1)–(3.2)–(3.10) globally asymptotically stable, where K is a constant matrix such that $A - KC$ is Hurwitz.

Proof: Let $\epsilon(k) = \xi(k) - x(k)$. Then the closed-loop system (3.1)–(3.2)–(3.10) can be expressed as

$$\left. \begin{aligned} \epsilon(k+1) &= (A - KC)\epsilon(k) + (g(\xi(k)) - g(x(k)))u_\alpha(k) \\ \xi(k+1) &= A\xi(k) + g(\xi(k))u_\alpha(k) - KC\epsilon(k) \end{aligned} \right\} \quad (3.11)$$

By assumption, there is a positive definite matrix R such that

$$(A - KC)^T R (A - KC) - R = -I$$

because $A - KC$ is Hurwitz. Let $W(\epsilon(k)) = \frac{1}{2}\epsilon^T(k)R\epsilon(k)$. Then

$$\begin{aligned} \Delta W(\epsilon(k)) &= W(\epsilon(k+1)) - W(\epsilon(k)) \\ &= -\frac{1}{2}\|\epsilon(k)\|^2 + \epsilon^T(k)(A - KC)^T \\ &\quad \times R(g(\xi(k)) - g(x(k)))u_\alpha(k) \\ &\quad + \frac{1}{2}u_\alpha^T(k)(g(\xi(k)) - g(x(k)))^T \\ &\quad \times R(g(\xi(k)) - g(x(k)))u_\alpha(k). \end{aligned}$$

From Lipschitzness of $g(\cdot)$ and $|u_\alpha(k)| \leq \alpha$, we deduce that

$$\begin{aligned} \Delta W(\epsilon(k)) &\leq -\frac{1}{2}\|\epsilon(k)\|^2 \\ &\quad \times (1 - 2\alpha N \|(A - KC)R\| - \alpha^2 N^2 \|R\|) \end{aligned}$$

where N is the Lipschitz constant associated with $g(\cdot)$. Obviously, it is possible to choose $\alpha > 0$ sufficient small so that for some $\theta, 0 < \theta < 1$

$$\begin{aligned} \Delta W(\epsilon(k)) &= \frac{1}{2}(\epsilon^T(k+1)R\epsilon(k+1) - \epsilon^T(k)R\epsilon(k)) \\ &\leq -\frac{\theta}{2}\epsilon^T(k)R\epsilon(k) \leq 0. \end{aligned}$$

This, in turn, implies

$$\begin{aligned} \epsilon^T(k)R\epsilon(k) &\leq (1 - \theta)\epsilon^T(k-1)R\epsilon(k-1) \\ &\leq \dots \leq (1 - \theta)^k \epsilon^T(0)R\epsilon(0). \end{aligned}$$

Thus

$$\|\epsilon(k)\| \leq M a^k \quad \text{for } k = 0, 1, \dots, M > 0 \text{ and } 0 < a < 1. \quad (3.12)$$

On the other hand, recall that by Theorem 3 [or inequality (3.9)], the following relationship

$$\|A\xi(k) + g(\xi(k))u_\alpha(k)\|_P^2 \leq \|\xi(k)\|_P^2, \quad k = 0, 1, \dots \quad (3.13)$$

is satisfied. Without loss of generality, Let $P = I$ in (3.13). Then, we deduce from (3.11) and (3.13) that

$$\begin{aligned} \|\xi(k+1)\| &\leq \|A\xi(k) + g(\xi(k))u_n(k)\| + \|K^*C\xi(k)\| \\ &\leq \|\xi(k)\| + M\alpha^k \|K^*C\| \leq \dots \\ &\leq \|\xi(0)\| + \frac{M\|K^*C\|}{1-\alpha}. \end{aligned} \quad (3.14)$$

From (3.12) and (3.14), we conclude that all trajectories of the closed-loop system (3.11) are bounded. Since $\lim_{k \rightarrow \infty} \epsilon(k) = 0$, the same arguments as the ones used in the proof of Theorem 2 show that every bounded trajectory of system (3.11) eventually converges to the invariant manifold characterized by the difference equation

$$\xi(k+1) = A\xi(k) + g(\xi(k))u_n(k)$$

which is globally asymptotically stable. Thus the proof is complete.

Finally, we turn our attention to the multi-input case. In this case, the input saturation technique developed in Theorem 3 cannot be applied directly to the multi-input discrete-time systems, because there exists a couple matrix $[I + \frac{1}{2}g^T(x)Pg(x)]^{-1}$ which appeared in the smooth state feedback control law. This is in sharp contrast to the case of continuous-time affine nonlinear systems (see Theorem 1). Nevertheless, using a modified state feedback control law, we can prove that it is possible to extend Theorems 3 and 4 to discrete-time MIMO nonlinear systems of the form (3.1)–(3.2).

Theorem 5: A multi-input discrete-time nonlinear system (3.1), which satisfies Hypothesis H2) and the assumption, $\Omega_s \cap S_s = \{0\}$, is always globally asymptotically stabilizable via smooth state feedback. In particular, a possible choice is

$$u = -\left[I + \frac{1}{2}g^T(x)Pg(x)\right]^{-1} g^T(x)PAx. \quad (3.15)$$

Moreover, system (3.1) can also be globally asymptotically stabilized by the bounded state feedback control law

$$\begin{aligned} u_n(x) &= -\alpha \left[I + \frac{1}{2}g^T(x)Pg(x)\right]^{-1} \\ &\quad \times \frac{g^T(x)PAx}{1 + \|g^T(x)PAx\|^2}, \text{ for any } 0 < \alpha < 1. \end{aligned} \quad (3.16)$$

Theorem 6: Under Assumptions H2) and $S \cap \Omega = \{0\}$, a discrete-time MIMO nonlinear system (3.1)–(3.2) can be globally asymptotically stabilized by the dynamic compensator

$$\begin{cases} \xi(k+1) = A\xi(k) + g(\xi(k))u(k) + K(y(k) - C\xi(k)) \\ u(k) = u_n(\xi(k)) \end{cases} \quad (3.17)$$

with any sufficient small $\alpha > 0$, provided that the pair (A, C) is observable and $g_i(x)$ is global Lipschitz for $1 \leq i \leq m$, where $u_n(\cdot)$ is given by (3.16) and K is such that $A - KC'$ is Hurwitz.

The proofs of Theorems 5 and 6 are strongly reminiscent of the proofs of Theorems 3 and 4 and therefore are left to the reader as an exercise.

Remark 1: As discussed in the last section, we note that discrete-time nonlinear systems (3.1)–(3.2) include bilinear systems as a particular case, a class of systems which has attracted considerable attention in the last two decades. Therefore, it is easy to see that all the global stabilization results developed in this section can be directly applied to discrete-time bilinear systems. Indeed, simply replacing $g_i(x)$ by $B_i x + N_i$ in Theorems 3–6 yields desired results for bilinear systems. It must be noted that Theorem 6 has provided a solution to the open problem raised in [11], namely the problem of global asymptotic stabilization via dynamic output feedback for MIMO bilinear systems.

IV. CONCLUSION

In this paper, we have presented smooth and bounded state feedback control schemes which globally asymptotically stabilize a class of continuous-time affine nonlinear systems whose free dynamics are Lyapunov stable. Based on an arbitrarily small bounded state feedback control law thus developed, we have been able to prove that a class of MIMO nonlinear systems which are "controllable" and "observable" is globally asymptotically stabilizable by a nonlinear Luenberger-observer-like based output feedback control law. The crucial point behind the development of a nonlinear enhancement of global separation principle is to use the technique of input saturation, which has attracted considerable attention in the literature recently in [13], [15], and [16]. The stabilization results established in continuous-time nonlinear systems have also been extended to their discrete-time counterparts. It must be pointed out that such generalizations are neither routine nor trivial. This certainly raises the value of Section III.

APPENDIX A

GLOBAL STABILIZATION OF AFFINE SYSTEMS BY BOUNDED STATE FEEDBACK

Consider an affine nonlinear control system

$$\Sigma: \dot{x} = f(x) + \sum_{i=1}^m g_i(x)u_i, \quad (A.1)$$

where $x \in \mathbb{R}^n$ is the state and $u_i \in \mathbb{R}$, $1 \leq i \leq m$, are control inputs, $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$, and $g_i: \mathbb{R}^n \rightarrow \mathbb{R}^n$, $1 \leq i \leq m$, are smooth mappings, with $f(0) = 0$. Suppose there exists a C^1 ($\nu \geq 1$) proper and positive definite function $V: \mathbb{R}^n \rightarrow \mathbb{R}$ such that

$$H) \quad L_f V(x) \leq 0 \quad \forall x \in \mathbb{R}^n$$

i.e., the unforced dynamics of (A.1) are Lyapunov stable. Then the following global stabilization result via bounded state feedback can be established.

Theorem A.1. An affine nonlinear control system (A.1) with property H) is always globally asymptotically stabilizable via arbitrarily small bounded state feedback provided that $\Omega_s \cap S_s = \{0\}$, where Ω_s and S_s denote the sets

$$\begin{aligned} \Omega_s &\triangleq \{x \in \mathbb{R}^n : L_k^i V(x) = 0, \quad 1 \leq k \leq r\} \\ S_s &\triangleq \{x \in \mathbb{R}^n : L_\tau^k V(x) = 0, \quad \forall \tau \in D, \quad 0 \leq k \leq r-1\} \end{aligned} \quad (A.2)$$

the distribution D is defined by

$$D = \text{span}\{ad_i^k g_i : 0 \leq k \leq n-1, 1 \leq i \leq m\}. \quad (A.3)$$

In particular, a typical bounded state feedback control law is given by

$$u_i^*(x) \triangleq -\text{sgn}(L_{g_i} V(x)) \min(|L_{g_i} V(x)|, \alpha), \quad 1 \leq i \leq m \quad (A.4)$$

where α is an arbitrarily positive real number.

Proof. Substituting the bounded state feedback control law (A.4) into system (A.1) yields the closed-loop system

$$\dot{x} = f(x) - \sum_{i=1}^m g_i(x) \text{sgn}(L_{g_i} V(x)) \min(|L_{g_i} V(x)|, \alpha). \quad (A.5)$$

Clearly, the equilibrium $x = 0$ of the closed-loop system is Lyapunov stable. As a matter of fact, a routine calculation proves that

$$\begin{aligned} \dot{V}(x) &= L_f V(x) - \sum_{i=1}^m L_{g_i} V(x) \text{sgn}(L_{g_i} V(x)) \\ &\quad \times \min(|L_{g_i} V(x)|, \alpha). \end{aligned}$$

Observe that $x \operatorname{sgn}(x) \equiv |x|$. With this in mind, it follows from Assumption H) that

$$\dot{V}(x) = L_f V(x) - \sum_{i=1}^m |L_{g_i} V(x)| \min(|L_{g_i} V(x)|, \alpha) \leq 0 \quad (\text{A.6})$$

Thus the claim follows.

To show that the equilibrium $x = 0$ of (A.5) is globally asymptotically stable, we let $\dot{V}(x) = 0$. This yields

$$L_f V(x) = 0 \quad \text{and} \quad L_{g_i} V(x) = 0, \quad 1 \leq i \leq m. \quad (\text{A.7})$$

Hence $u_i^*(x) = 0$ for $1 \leq i \leq m$. Let $x(t, x_0)$ be the trajectory of the free dynamics $\dot{x} = f(x) \forall t \geq 0$ starting from $x(0) = x_0$. Let r^0 denote its ω -limit set. Since $L_f V(x) \leq 0$, r^0 is nonempty, compact and invariant. Therefore for any initial $\bar{x} \in r^0$, the corresponding trajectory $x(t, \bar{x})$ of $\dot{x} = f(x)$ stays in r^0 forever. Note that $V(x(t))$ is positive definite and nonincreasing along every trajectory of the free dynamics. Thus $\lim_{t \rightarrow \infty} V(x(t)) = \beta \geq 0$. By continuity of V , $V(\bar{x}) = \beta$ for every $\bar{x} = \lim_{j \rightarrow \infty} x(t_j, x_0)$ in r^0 . In other words, $V(x)$ is constant over r^0 . In particular, $\forall t \geq 0$

$$L_f^k V(x(t, \bar{x})) = 0 \quad \forall \bar{x} \in r^0, \quad 1 \leq k \leq r. \quad (\text{A.8})$$

On the other hand, since $L_{g_i} V(x)$ vanishes $\forall x \in \{x: \dot{V}(x) = 0\}$, so does $L_{g_i} V(\bar{x}) \forall \bar{x} \in r^0$. Thus $L_{g_i} V(x(t, \bar{x})) = 0 \forall t \geq 0, 1 \leq i \leq m$. This, together with (A.7), implies that for $\tau = [f, g_i]$

$$\begin{aligned} L_\tau V(x(t, \bar{x})) &= L_f L_{g_i} V(x(t, \bar{x})) - L_{g_i} L_f V(x(t, \bar{x})) \\ &= 0, \quad 1 \leq i \leq m. \end{aligned}$$

By induction, it is straightforward to prove that

$$L_f^k L_\tau V(x(t, \bar{x})) = 0 \quad \forall \bar{x} \in D_\Sigma \text{ and } 0 \leq k \leq r-1. \quad (\text{A.9})$$

According to the arguments above, we conclude from (A.8)–(A.9) and LaSalle's Invariant Principle that all trajectories of the closed-loop system (A.6) eventually approach the largest invariant set I of $\{x \in \mathbb{R}^n: \dot{V} = 0\}$, which is contained in $S_\Sigma \cap \Omega_\Sigma$. By assumption, $\Omega_\Sigma \cap S_\Sigma = \{0\}$. This implies $I = \{x \in \mathbb{R}^n: \dot{V} = 0\} = \{0\}$, thus completing the proof of global asymptotic stability.

Let $\alpha = +\infty$. We deduce immediately from Theorem A.1 the following result which is a slight extension of Theorem 1 proposed in [12].

Corollary A.2: An affine nonlinear system (A.1) whose free dynamics are Lyapunov stable can always be globally asymptotically stabilizable by smooth state feedback if $\Omega_\Sigma \cap S_\Sigma = \{0\}$. In particular, a smooth state feedback control law is given by

$$u = -(L_{g_m} V(x))^T = -[L_{g_1} V(x), \dots, L_{g_m} V(x)]^T. \quad (\text{A.10})$$

APPENDIX B

A BOUNDEDNESS CRITERION FOR PERTURBED SYSTEMS

Theorem B.1: Consider a perturbed system described by differential equations of the form

$$\Sigma_p: \dot{x} = f(x) + d(t). \quad (\text{B.1})$$

Suppose there exists a C^1 function $V: \mathbb{R}^n \rightarrow \mathbb{R}$, with $V(0) = 0$, such that

$$\text{A1) } V(x) \geq a_1 \|x\|^2, \quad \left\| \frac{\partial V}{\partial x} \right\| \leq a_2 \|x\| \text{ and } L_f V(x) \leq 0.$$

Assume that the disturbance $d: \mathbb{R} \rightarrow \mathbb{R}^n$ is piecewise continuous and satisfies

$$\text{A2) } \int_0^\infty \|d(t)\| dt \leq a_3 < +\infty$$

where $a_i, 1 \leq i \leq 3$, are positive real constants. Then all the trajectories of system (B.1) are bounded.

Proof: By Assumption A1)

$$\dot{V} = L_f V(x) + \frac{\partial V}{\partial x} d(t) \leq \frac{\partial V}{\partial x} d(t).$$

Hence

$$\begin{aligned} V(x(t)) &\leq V(x_0) + \int_0^t \frac{\partial V}{\partial x} d(\tau) d\tau \\ &\leq V(x_0) + \int_0^t \left\| \frac{\partial V}{\partial x} \right\| \|d(\tau)\| d\tau. \end{aligned}$$

Using Assumption A1) again, we have

$$a_1 \|x(t)\|^2 \leq V(x_0) + a_2 \int_0^t \|x(\tau)\| \|d(\tau)\| d\tau. \quad (\text{B.2})$$

Thus

$$\begin{aligned} a_1 \|x(t)\| &\leq a_1 \|x(t)\|^2 + a_1 \\ &\leq a_1 + V(x_0) + a_2 \int_0^t \|x(\tau)\| \|d(\tau)\| d\tau. \end{aligned}$$

This, in turn, implies

$$\|x(t)\| \leq \alpha_1 + \alpha_2 \int_0^t \|x(\tau)\| \|d(\tau)\| d\tau$$

for some $\alpha_i \geq 0, i = 1, 2$. From the Bellman–Gronwall inequality, we obtain the estimate

$$\|x(t)\| \leq \alpha_1 e^{\alpha_2 \int_0^t \|d(\tau)\| d\tau} \quad \forall t \geq 0. \quad (\text{B.3})$$

Because of A2), Theorem B.1 follows from (B.4).

Remark. A similar argument shows that Theorem B.1 remains true if A1) is replaced by the following assumption

$$\text{A) } V(x) \geq a_1 \|x\|^{q+1}, \quad \left\| \frac{\partial V}{\partial x} \right\| \leq a_2 \|x\|^q$$

and

$$L_f V(x) \leq 0, \text{ for any positive integer } q.$$

REFERENCES

- [1] C. I. Byrnes, A. Isidori, and J. C. Willems, "Passivity, feedback equivalence, and global stabilization of minimum phase nonlinear systems," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1228–1240, 1991.
- [2] C. I. Byrnes and W. Lin, "Losslessness, feedback equivalence and the global stabilization of discrete-time nonlinear systems," *IEEE Trans. Automat. Contr.*, vol. 39, no. 1, pp. 83–98, 1994.
- [3] —, "Stabilization of discrete-time nonlinear systems by smooth state feedback," *Syst. Contr. Lett.*, vol. 21, pp. 255–263, 1993.
- [4] W. Lin and C. I. Byrnes, "Passivity and absolute stabilization of a class of discrete-time nonlinear systems," *Automatica*, vol. 31, no. 2, pp. 263–268, 1995.
- [5] J. Carr, *Application of Centre Manifold Theory*. New York: Springer-Verlag, 1981.
- [6] J. P. Gauthier and G. Bornard, "Stabilization des systèmes nonlinéaires, Outillage mathématiques pour l'automatique," I. D. Landau, Ed., *C.N.R.S.* 1981, pp. 307–324.
- [7] J. P. Gauthier and I. Kupka, "A separation principle for bilinear systems with dissipative drift," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1970–1974, 1992.
- [8] V. Jurdjevic and J. P. Quinn, "Controllability and stability," *J. Diff. Equations*, vol. 28, pp. 381–389, 1978.
- [9] N. Kalouptsidis and J. Tsinnas, "Stability improvement of nonlinear systems by feedback," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 364–367, 1984.

- [10] W. Lin, "Input saturation and global stabilization by output feedback for affine nonlinear systems," in *Proc. 33rd IEEE CDC*, 1994, pp. 1323-1328.
- [11] W. Lin and C. I. Byrnes, "KYP Lemma, State feedback and dynamic output feedback in discrete-time bilinear systems," *Syst. Contr. Lett.*, vol. 23, pp. 127-136, 1994.
- [12] K. K. Lee and A. Arapostathis, "Remarks on smooth feedback stabilization of nonlinear systems," *Syst. Contr. Lett.*, vol. 10, pp. 41-44, 1988.
- [13] Z. Lin, "Global and semi-global control problems for linear systems subject to input saturation and minimum-phase input-output linearizable systems," Ph.D. dissertation, Washington State Univ., Pullman, Dec. 1993.
- [14] H. Nijmeijer and A. J. Van der Schaft, *Nonlinear Dynamic Control Systems*. New York: Springer-Verlag, 1990.
- [15] E. D. Sontag and H. J. Sussmann, "Nonlinear output feedback design for linear systems with saturating controls," in *Proc. 29th IEEE CDC*, 1990, pp. 3414-3416.
- [16] A. R. Teel, "Global stabilization and restricted tracking for multiple integrators with bounded controls," *Syst. Contr. Lett.*, vol. 18, pp. 165-171, 1992.
- [17] M. Vidyasagar, "On the stabilization of nonlinear systems using state detection," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 504-509, 1980.

Optimal Control for Systems with Deterministic Production Cycles

Jian-Qiang Hu and Dong Xiang

Abstract—In this paper we consider a failure prone production system with deterministic production cycles. The objective is to find the optimal production rate to minimize long-run average cost. We first show that the optimal control policy belongs to a type of switching curve policy which has a special structural property and can be characterized by a single parameter. We then establish a relationship between the surplus process of the system under the optimal control policy and the workload process of a $D/G/1$ queue, based on which the existing results in queueing theory which can be applied to obtain the steady-state probability distribution of the surplus process under the optimal control policy.

1. INTRODUCTION

Consider a production system which has a single machine and produces a single part-type. The system meets a constant demand rate d and backlog is allowed. The machine has two states: up and down, which are denoted by one and zero, respectively. The produced parts are represented by a fluid flow. When the machine is up, it can produce at any rate between zero and a maximum value r . Denote the production surplus at time t by X_t ; a positive value of X_t represents inventory while a negative value represents backlog. Let $\alpha_t \in \{0, 1\}$ be the state of the machine at time t and u_t be the controlled production rate of the machine at time t under a control policy π . X_t can then be characterized by the following differential

equation

$$\frac{dX_t}{dt} = u_t - d \quad (1)$$

where $0 \leq u_t \leq \alpha_t r$. For each control policy π , we use J_π to denote its long-run average expected cost

$$J_\pi = \lim_{T \rightarrow \infty} \frac{1}{T} E \int_0^T (C^+ X_t^+ + C^- X_t^-) dt \quad (2)$$

where $X_t^+ = \max(X_t, 0)$, $X_t^- = \max(-X_t, 0)$ and C^+ , C^- are two nonnegative constants. The goal is to find a control policy π^* to minimize J_{π^*} .

A few variations of the above problem (e.g., systems with multiple machines and/or multiple part types) have been studied by many researchers (see [1], [2] and references therein). The key assumption used in almost all previous work, with a few exceptions [3]–[9], is that the machine state process $\{\alpha_t\}$ is a Markov process, i.e., the machine up and down times are exponentially distributed. Under such an assumption, the system can be modeled as a system with jump Markov disturbances based on Rishel's formulation ([3]), and it can be shown [10] that the optimal control is a hedging point policy. The Markov assumption, however, is not very realistic in many applications. So far, only limited work has been done for general systems.

In this paper we consider the one machine and one part-type system in which the machine has deterministic up time and general down time. A production system with deterministic up time, which is similar to ours, is also studied by Meyer *et al.* [11]. The system we consider here can be ideally used to model production systems in which production is only interrupted after (approximately) a fixed amount of time since it starts. For example, consider a production system in which preventive maintenance is performed after the system is in operation for a fixed amount of time. If the system rarely fails due to its scheduled preventive maintenance, then its up time is approximately deterministic. Another example is an inventory system which replenishes its inventory continuously (with no interruption) but only receives demands (orders) periodically. In such a system each period corresponds to one up time and an order (a random variable) received at the end of each period corresponds to total demands accumulated during one down time. It is also worth pointing out that the periodic review system with limited capacity studied in inventory theory ([12]) is very similar to our system and their relationship is investigated in a recent paper by Fu and Hu [14]. As we shall see, the well-known hedging point policy is no longer optimal for our system. Our analysis shows, however, that its optimal control policy belongs to a class of so-called switching curve policies whose switching curves have a very simple structure and can be characterized by a single parameter. Furthermore, we show that the surplus process of the system operated under a switching curve policy is related to the workload process of a $D/G/1$ queue, a result similar to the one established in [15] for the one machine and one part-type system under a hedging point policy (also see [16]–[19] for similar results on the relationship of fluid models to the workload process of queueing systems). This enables us to use the existing results in queueing theory to obtain the steady-state distribution function of the surplus process for the class of parameterized switching curve policies to which the optimal control belongs.

If $\{X_t; t \geq 0\}$ has steady-state probability distribution function $F(x)$ under a control policy π , then the average expected cost J_π

Manuscript received October 13, 1994. This work was supported in part by National Science Foundation Grants DDM-8914277, EID-9212122, and DDM-9212368.

The authors are with the Department of Manufacturing Engineering, Boston University, Boston, MA 02215 USA.

IEEE Log Number 9408787.

can be computed as the expected cost with respect to $F(x)$, i.e.,

$$J_\pi = \int_{-\infty}^{\infty} (C^+ x^+ + C^- x^-) dF(x). \quad (3)$$

Therefore, with the steady-state probability we obtained we can finally convert the original optimal control problem into an optimization problem with a single parameter, which is much simpler and easier to solve.

The rest of this paper is organized as follows. In Section II, we first establish a structural property for the optimal control. Then, we discuss the switching curve policies and show that the optimal switching curve can be characterized by a single parameter. In Section III, we show how the surplus process of the system under a switching curve policy is related to the workload process of a $D/G/1$ queue. This relationship is used in Section IV to obtain the steady-state probability distribution of the surplus process for the system under the class of switching curve policies to which the optimal switching curve policy belongs. A simple example with exponential machine down times is provided in Section V.

II. THE OPTIMAL CONTROL POLICY

In this section, we first establish a structural property for the optimal control based on sample path analysis. Then we discuss a class of so-called switching curve policies. Using the structural property established, we can show that when restricting ourselves to the switching curve policies, the switching curve for the optimal control belongs to a class of switching curves which can be characterized by a single parameter.

Intuitively, the structural property of the optimal control we are about to establish can be interpreted as follows: to minimize the surplus cost, one needs to maintain a nonnegative surplus at the end of each machine up period to anticipate future capacity shortages brought by machine failures while keeping the surplus level close to zero as much as possible. Most of our derivations are quite straightforward. We hope that the accompanying figures greatly aid in the understanding of the derivations.

Suppose π^* is the optimal control policy. Given the machine state process $\{\alpha_t\}$, let $\{u_t^*\}$ be the optimal control process generated under the optimal control policy π^* , and $\{X_t^*\}$ be the corresponding surplus process determined by (1) with u_t being replaced by u_t^* .

Lemma 1: There exists an optimal control $\{u_t^*\}$ that satisfies the following condition

$$u_t^* \begin{cases} = r & \text{if } X_t^* < 0, \alpha_t = 1; \\ \geq d & \text{if } X_t^* = 0, \alpha_t = 1. \end{cases} \quad (4)$$

Condition (4) can be explained as follows: When the surplus is negative the production rate should be set at the maximum rate r so that the surplus can be brought back to zero level as quickly as possible; when the surplus is at zero level we should keep it at a nonnegative level as long as we can. In other words, we should always prevent the surplus from being negative. Based on Lemma 1, we shall henceforth assume the optimal control we consider satisfies (4).

Proof: If $\{u_t^*\}$ does not satisfy (4), we can construct another optimal control based on $\{u_t^*\}$ which satisfies (4). Suppose $\alpha_{t_0} = 1$. Let us first consider the case $X_{t_0}^* < 0$. We denote the first hitting time to 0 from $X_{t_0}^*$ by

$$t_0^b = \inf\{t: X_t^* = 0, t > t_0\}.$$

By definition, we have $X_t^* < 0$ for $t \in [t_0, t_0^b)$. We construct

$$u_t = \begin{cases} r\alpha_t, & \text{if } t \in [t_0, t_0^b) \text{ and } X_t < 0; \\ d\alpha_t, & \text{if } t \in [t_0, t_0^b) \text{ and } X_t = 0; \\ u_t^*, & \text{otherwise} \end{cases} \quad (5)$$

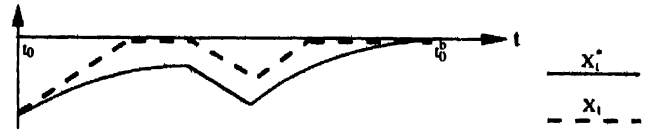


Fig. 1. Comparison between X_t^* and X_t during $[t_0, t_0^b]$ for the case $X_{t_0}^* < 0$.

where $\{X_t\}$ is the surplus process associated with $\{u_t\}$. Under the control $\{u_t\}$, the machine produces at the maximum rate r (if possible) until the surplus level reaches zero and then remains at zero (if possible) until t_0^b . Therefore, we have $u_{t_0} = r$ (in fact, $\{u_t\}$ satisfies (4) at least during the time interval $[t_0, t_0^b]$) and

$$X_t \begin{cases} \geq X_t^*, & \text{if } t \in [t_0, t_0^b]; \\ = X_t^*, & \text{otherwise} \end{cases}$$

(see Fig. 1), from which we immediately obtain

$$\begin{aligned} & \int_0^T (C^+ X_t^+ + C^- X_t^-) dt \\ & \leq \int_0^T (C^+ (X_t^*)^+ + C^- (X_t^*)^-) dt \quad \text{for any } T \geq 0. \end{aligned} \quad (6)$$

That is to say $\{u_t\}$ is also an optimal control process.

For the case $X_{t_0}^* = 0$, we define

$$t_0^b = \inf\{t: X_t^* > 0, t > t_0\}.$$

If $t_0^b = t_0$, we then have $dX_t^*/dt \geq 0$, which implies $u_t^* \geq d$. If $t_0^b > t_0$, we have $X_t^* \leq 0$ for $t \in (t_0, t_0^b)$, in which case we construct $\{u_t\}$ based on (5) with t_0^b being replaced by t_0^b . The rest of the proof remains the same. \square

Lemma 2: There always exists an optimal control $\{u_t^*\}$ such that it only takes three values 0, d , and r , and its value increases with respect to the age of the machine up time when the surplus level becomes nonnegative.

Proof: Denote the (deterministic) machine up time by D . We consider a time interval $[t_1, t_2]$ during which the machine is up, where t_1 is the epoch at which the machine up period is initiated and $t_2 = t_1 + D$ (i.e., t_2 is the time epoch at which the machine up period ends). Unless otherwise stated, in what follows we shall focus on the interval $[t_1, t_2]$. Since $\{u_t^*\}$ satisfies (4), it is not difficult for us to verify that

- i) If $X_{t_1}^* < 0$, then $X_{t_2}^* \geq X_{t_1}^*$;
- ii) If $X_t^* \geq 0$ for some $t \in [t_1, t_2]$, then $X^*(\tau) \geq 0$ for all $\tau \in [t, t_2]$.

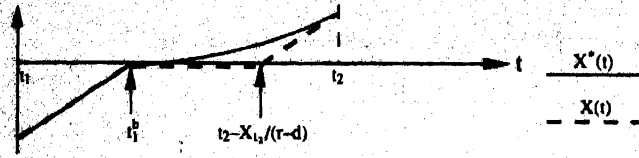
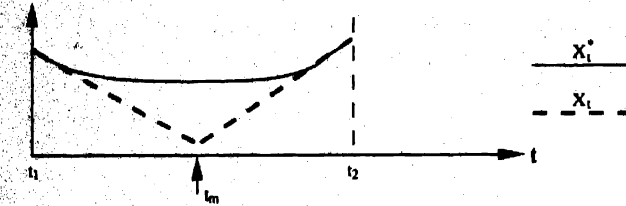
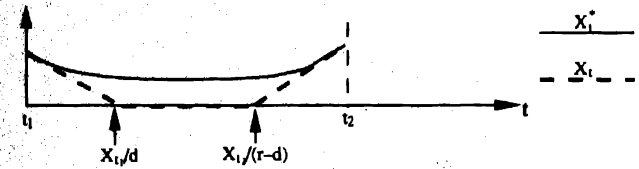
Therefore, we have the following three cases:

- 1) $X_{t_1}^* < 0$ and $X_{t_2}^* < 0$. In this case $X_t^* < 0$ for all $t \in [t_1, t_2]$, which implies $u_t^* = r$ for all $t \in [t_1, t_2]$.
- 2) $X_{t_1}^* < 0$ and $X_{t_2}^* \geq 0$. Let t_1^b be the first hitting time to zero from $X_{t_1}^*$. It immediately follows that $u_t^* = r$ for $t \in [t_1, t_1^b)$ since $X_t^* < 0$ for $t \in [t_1, t_1^b)$. Also it is clear that $X_t^* \geq 0$ for $t \in [t_1^b, t_2]$. We now construct another control $\{u_t\}$ based on $\{u_t^*\}$ as follows

$$u_t = \begin{cases} d, & \text{if } t \in [t_1^b, t_2 - X_{t_2}^*/(r-d)); \\ r, & \text{if } t \in [t_2 - X_{t_2}^*/(r-d), t_2]; \\ u_t^*, & \text{otherwise.} \end{cases}$$

We can easily verify that the surplus process $\{X_t\}$ associated with the control $\{u_t\}$ satisfies

$$X_t \begin{cases} \leq X_t^*, & \text{if } t \in [t_2 - X_{t_2}^*/(r-d), t_2]; \\ = X_t^*, & \text{otherwise} \end{cases}$$

Fig. 2. Comparison between X_t^* and X_t during $[t_1, t_2]$ for Case 2.Fig. 3. Comparison between X_t^* and X_t during $[t_1, t_2]$ for Case 3.1.Fig. 4. Comparison between X_t^* and X_t during $[t_1, t_2]$ for Case 3.2.

(see Fig. 2), which immediately leads to (6). Hence, $\{u_t\}$ is also an optimal control.

- 3) $X_{t_1}^* \geq 0$ and $X_{t_2}^* \geq 0$. In this case we have $X_t \geq 0$ for all $t \in [t_1, t_2]$. There are two subcases to consider:

- 3.1). $X_{t_2}^*/(r-d) + X_{t_1}^*/d \geq D$. We construct the following control $\{u_t\}$ based on $\{u_t^*\}$

$$u_t = \begin{cases} 0, & \text{if } t \in [t_1, t_1 + t_m]; \\ r, & \text{if } t \in [t_1 + t_m, t_2]; \\ u_t^*, & \text{otherwise} \end{cases}$$

where $t_m = (X_{t_1}^* + (r-d)D - X_{t_2}^*)/r$. Similar to Case 2 we can then show that the surplus process $\{X_t\}$ associated with the control $\{u_t\}$ satisfies

$$X_t \begin{cases} \leq X_t^*, & \text{if } t \in [t_1, t_2]; \\ = X_t^*, & \text{otherwise} \end{cases}$$

(see Fig. 3), from which (6) follows immediately. Therefore, $\{u_t\}$ is also an optimal control.

- 3.2). $X_{t_2}^*/(r-d) + X_{t_1}^*/d < D$. If we construct the following control $\{u_t\}$

$$u_t = \begin{cases} 0, & \text{if } t \in [t_1, t_1 + X_{t_1}^*/d]; \\ d, & \text{if } t \in [t_1 + X_{t_1}^*/d, t_2 - X_{t_2}^*/(r-d)]; \\ r, & \text{if } t \in [t_2 - X_{t_2}^*/(r-d), t_2]; \\ u_t^*, & \text{otherwise} \end{cases}$$

then the same result as that of Case 3.1 can be obtained (see Fig. 4). \square

We now consider a class of so-called switching curve policies.

Definition 1: A switching curve policy π is defined by

$$u_t = \pi(x_t, a_t) = \begin{cases} r\alpha_t & \text{if } x_t < S(a_t); \\ (d + S'(a)|_{a=a_t})\alpha_t & \text{if } x_t = S(a_t); \\ 0 & \text{if } x_t > S(a_t) \end{cases}$$

where $S: [0, D] \rightarrow \mathbb{R}$ is piecewise differentiable and its derivative $S'(a)$ (whenever it exists) satisfies $-d \leq S'(a) \leq r-d$. We call $S(\cdot)$ the switching curve.

The following result follows immediately from the proof of Lemma 2.

Lemma 3: The optimal switching curve is given by

$$S(a) = \begin{cases} 0 & \text{if } 0 \leq a \leq D - z/(r-d), \\ z - (D-a)(r-d) & \text{if } D - z/(r-d) < a \leq D. \end{cases} \quad (7)$$

In (7) z is the inventory level at which the surplus needs to be maintained at the end of the machine up period to hedge against future capacity shortages brought by machine failures. Therefore, it plays a similar role as the hedging point in a hedging point policy. There is a significant difference, however, between the hedging point policy and the control policy defined by (7). Under the hedging point policy, the machine produces at its maximum rate until the surplus reaches the hedging level (i.e., the surplus is brought to the hedging level as fast as it could be). Under (7) one only has to make sure that the surplus is brought to the level z at the end of each machine up period and before that the surplus level should be kept as close as to zero (therefore minimizing the surplus cost). Obviously, this is because the machine up time is now deterministic, and we know exactly when the machine will fail so that the machine only needs to produce parts at its maximum rate toward the end of each machine up period to guarantee that the surplus level will eventually reach the level z before the machine fails.

In the remainder of this paper, we restrict ourselves to the switching curve policies. We shall show how to obtain the steady-state probability distribution of the surplus process for the system operated under the switching curve policy (7).

III. QUEUEING EQUIVALENCE

In this section we study the surplus process of the system under a switching curve policy. We shall show that the surplus levels at instances when the machine is up and down are "equivalent" to the system and waiting times of jobs of a $D/G/1$ queue, a result which is similar to the one established in Hu and Xiang [15] for the one machine and one part-type system under the hedging point policy. This queueing equivalent result enables us to use the existing results for single-server queues to obtain the steady-state probability distribution for the surpluses at these instances, based on which we can then derive the steady-state probability distribution of the surplus process for the system under the switching curve policy (7) using the level crossing technique.

Denote $S(D)$ by z . For the sake of convenience, we assume $X_0 = z$ and $a_0 = D$ (i.e., $\alpha_{0-} = 1$ and $\alpha_{0+} = 0$). Denote by $t_{d,n}$ the length of the n th down time and by $t_{u,n}$ the length of the n th up time. Therefore, $t_{d,n}$'s are i.i.d. random variables, say, with probability distribution function $G(x)$ and density function $g(x)$, and $t_{u,n}$'s are equal to D . Let $T_{d,n}$ be the epoch at which the n th machine down time starts and $T_{u,n}$ be the epoch at which the n th machine up time starts, i.e.,

$$T_{d,n} = \sum_{i=1}^{n-1} (t_{d,i} + t_{u,i}) \quad \text{and} \quad T_{u,n} = T_{d,n} + t_{d,n}. \quad (8)$$

For simplicity, we denote

$$X_{d,n} \triangleq X_{T_{d,n}} \quad \text{and} \quad X_{u,n} \triangleq X_{T_{u,n}}.$$

Let $Y_{d,n} = z - X_{d,n}$ and $Y_{u,n} = z - X_{u,n}$. Then applying the same method used in [15], we can show that $Y_{d,n}$ and $Y_{u,n}$ correspond to the waiting time and the system time of job n in a $GI/G/1$

queue with service times $\{t_{d,n}d: n = 1, 2, \dots\}$ and interarrival times $\{t_{u,n}(r-d): n = 1, 2, \dots\}$. Since $t_{u,n}(r-d)$ is equal to a constant $D(r-d)$, the $GI/G/1$ queue is in fact a $D/G/1$ queue. Therefore, we can apply various methods developed for single server queues in the literature to obtain the steady-state probability distribution functions of $Y_{d,n}$ and $Y_{u,n}$ (hence $X_{d,n}$ and $X_{u,n}$) (e.g., see [20]). In the following sections, we shall simply assume that the steady-state probability distributions of $X_{d,n}$ and $X_{u,n}$ are available.

IV. DISTRIBUTION FUNCTION

In this section, we consider the switching curve policy (7). We use the level crossing technique to derive the steady-state distribution function of the surplus process under (7). We should point out, however, that the technique can also be used to obtain the steady-state distribution function of the surplus process under a general switching curve policy though the derivation becomes a bit more involved.

Denote the steady-state distribution function of $X_{u,n}$ by $F_{X_{u,n}}(x)$, and its density function by $f_{X_{u,n}}(x)$. Define

$D_t(x)$ = the number of x downcrossings of the process $\{X_t\}$ during $[0, t]$;

$U_t(x)$ = the number of x -upcrossings of the process $\{X_t\}$ during $[0, t]$.

Notice that the accumulated time that the process $\{X_t\}$ takes value between $[x, x + \delta x]$ during $[0, t]$ is equal to

$$\frac{U_t(x)\delta x}{t-d} + \frac{D_t(x)\delta x}{d}.$$

Therefore, the steady-state probability density function of $\{X_t\}$ is given by

$$f_X(x) = \lim_{t \rightarrow \infty} \frac{1}{t} \left(U_t(x) + \frac{D_t(x)}{d} \right) \quad \text{for } x \leq z \text{ and } x \neq 0. \quad (9)$$

(The interested reader is referred to [16] for a rigorous derivation of (13).) Note that

$$\lim_{t \rightarrow \infty} \frac{U_t(x)}{t} = \lim_{t \rightarrow \infty} \frac{D_t(x)}{t}$$

it then follows from (13) that

$$\begin{aligned} f_X(x) &= \frac{r}{d(r-d)} \lim_{t \rightarrow \infty} \frac{U_t(x)}{t} \\ &= \frac{r}{d(r-d)} \lim_{n \rightarrow \infty} \frac{U_{t_{u,n+1}}(x)}{T_{u,n+1}} \quad \text{for } x \leq z \text{ and } x \neq 0. \end{aligned} \quad (10)$$

(Recall $T_{u,n+1}$ is the epoch at which the $(n+1)$ th machine up time initiated.) To calculate $\lim_{n \rightarrow \infty} U_{t_{u,n+1}}(x)/T_{u,n+1}$, we consider an down and up cycle $[T_{u,n}, T_{u,n+1}]$. First, it is obvious that the process $\{X_t\}$ can have at most one x -upcrossing during $[T_{u,n}, T_{u,n+1}]$ (in fact during $[T_{u,n}, T_{d,n+1}]$). Define

$$c(x) = \begin{cases} D - \frac{x}{r-d}, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0. \end{cases}$$

(See Fig. 5.) Then $\{X_t\}$ has one x -upcrossing during $[T_{u,n}, T_{d,n+1}]$ if and only if $x - (r-d)D < X_{u,n} \leq x + dc(x)$. Therefore

$$U_{t_{u,n+1}}(x) = \sum_{i=1}^n \mathbf{1}(x - (r-d)D < X_{u,i} \leq x + dc(x)).$$

It follows that

$$f_X(x) = \frac{r}{d(r-d)} \lim_{n \rightarrow \infty} \frac{n}{T_{u,n+1}} \frac{1}{n}$$

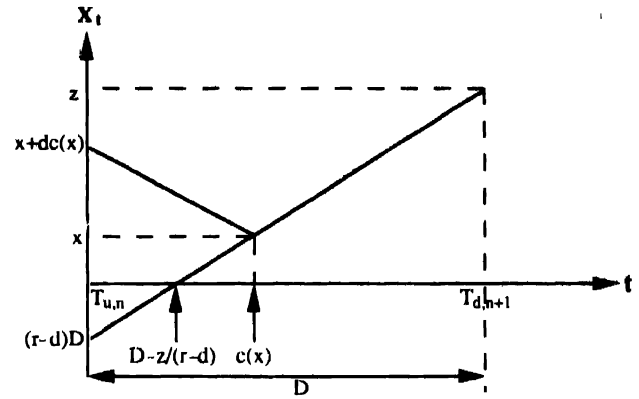


Fig. 5. Definition of $c(x)$.

$$\begin{aligned} & \times \sum_{i=1}^n \mathbf{1}(x - (r-d)D < X_{u,i} \leq x + dc(x)) \\ & = b[F_{X_{u,n}}(x + dc(x)) - F_{X_{u,n}}(x - (r-d)D)] \\ & \quad \text{for } x \leq z \text{ and } x \neq 0 \end{aligned} \quad (11)$$

where $b = r/((D+R)d(r-d))$ and R is the average machine down time. Taking the normalization factor into consideration, we have

$$F_X(0) = 1 - \int_{-\infty}^z f_X(x) dx. \quad (12)$$

Let J be the run-long average expected cost associated with the switching curve policy [7], then

$$J = \int_{-\infty}^z (C^+ x^+ + C^- x^-) f_X(x) dx.$$

Therefore, the optimal switching curve (i.e., the optimal value of z) can be obtained by solving the following optimization problem: $\min J$.

V. AN EXAMPLE

In this section, we study a simple example in which the machine down time is exponentially distributed with mean R and rate $\mu (= 1/R)$. We will use the results obtained in the previous section to calculate $f_X(x)$ and J . Since now the down time is exponentially distributed, the corresponding $D/G/1$ queue becomes a $D/M/1$ with the mean interarrival time $(r-d)D$ and the mean service time Rd . Hence we have ([20])

$$F_{X_{u,n}}(x) = e^{-(1-\sigma)(-x)\mu/d} \quad \text{for } x \leq z$$

where σ is the unique solution to the following equation in the range $0 \leq \sigma < 1$

$$\sigma = e^{-(1-d)D(1-\sigma)\mu/d}.$$

Based on [15] we obtain

$$\begin{aligned} f_X(x) &= b(e^{-(1-\sigma)\mu(-x+(r-d)D)/d} \\ & \quad - e^{-(1-\sigma)\mu(-x+(r-d)D)/d}) \\ & \quad \text{for } x \leq z \text{ and } x \neq 0. \end{aligned} \quad (13)$$

When $z < d(r-d)D/r$, (13) gives

$$\begin{aligned} f_X(x) &= \\ & \begin{cases} b(1 - e^{-(1-\sigma)\mu(-x+(r-d)D)/d}) & \text{if } 0 < x \leq z, \\ b e^{-(1-\sigma)\mu(-x)/d} (1 - e^{-(1-\sigma)\mu(r-d)D/d}) & \text{if } x < 0 \end{cases} \end{aligned}$$

and

$$J_s = \frac{C^+ b \omega^2}{2} - C^+ b e^{-\mu(1-\sigma)(1-d)D/r} \left[\frac{d}{\mu(1-\sigma)} - \frac{d^2}{\mu^2(1-\sigma)^2} + \frac{d^2}{\mu^2(1-\sigma)^2} e^{-\mu(1-\sigma)/r} \right] + \frac{C^+ b d^2}{\mu^2(1-\sigma)^2} e^{-\mu(1-\sigma)/r} (1 - e^{-\mu(1-\sigma)(1-d)D/r})$$

When $\omega \geq d(1-d)D/r$, (13) gives

$$f_s(x) = \begin{cases} b(1 - e^{-(1-\sigma)\mu(x + (1-d)D/r)}) & \text{if } x - d(1-d)D/r < x < \omega, \\ b(e^{-(1-\sigma)\mu(x - (1-d)D/r)} - e^{-(1-\sigma)\mu(\omega - (1-d)D/r)}) & \text{if } 0 < x \leq -d(1-d)D/r, \\ b e^{-(1-\sigma)\mu(x)/r} (1 - e^{-(1-\sigma)\mu(\omega - (1-d)D/r)}) & \text{if } x < 0 \end{cases}$$

and

$$J = \frac{bC^+ d}{\mu(1-\sigma)} \times \left(\frac{1-d}{r} (D(1-\sigma)\mu + 1) - e^{-\mu(1-\sigma)(1-d)D/r} \right) + \frac{bC^+ d^2}{\mu^2(1-\sigma)^2} (1 - e^{-\mu(1-\sigma)/r}) e^{-(1-\sigma)\mu(\omega - (1-d)D/r)} + \frac{bC^+ d^2}{\mu^2(1-\sigma)^2} e^{-\mu(1-\sigma)/r} (1 - e^{-(1-\sigma)\mu(\omega - (1-d)D/r)}) - \frac{bC^+ d^2(1-d)^2 D}{2r^2} \left(\frac{2}{(1-\sigma)\mu} + D \right) + \frac{bC^+ d^2(1-d)^2}{\mu^2(1-\sigma)^2 r^2} \left(e^{-(1-\sigma)\mu(\omega - (1-d)D/r)} - 1 \right)$$

REFERENCES

- [1] S. Gershwin, *Manufacturing Systems Engineering*, Englewood Cliffs, NJ: Prentice Hall, 1993.
- [2] S. P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Boston, MA: Birkhauser, 1994.
- [3] R. Rishel, "Dynamic programming and minimum principles for systems with Jump Markov distributions," *SIAM J. Contr. Optim.*, vol. 13, pp. 338-371, 1975.
- [4] J. Q. Hu and D. Xiang, "Structural properties of optimal production controllers in failure prone manufacturing systems," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 640-642, 1994.
- [5] J. Q. Hu and D. Xiang, "Monotonicity of optimal flow control for failure prone production systems," *J. Optim. Theory Application*, vol. 86, 1995, to appear.
- [6] J. Q. Hu, P. Vakili, and G. X. Yu, "Optimality of hedging point policies in the production control of failure prone manufacturing systems," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 1875-1880, 1994.
- [7] G. Liberopoulos, "Flow control of failure prone manufacturing systems: Control design theory and applications," Ph.D. dissertation, Manuf. Eng. Dept., Boston Univ., MA, 1992.
- [8] G. Liberopoulos and M. Caramanis, "Production control of manufacturing systems with production rates dependent failure rate," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 889-895, 1994.
- [9] F. S. Tu, D. P. Song, and S. X. C. Lou, "Preventive hedging point control policy," Preprint, 1992.
- [10] T. Bielecki and P. R. Kumar, "Optimality of zero inventory policies for unreliable manufacturing systems," *Op. Res.*, vol. 36, pp. 532-541, 1988.
- [11] R. R. Meyer, M. H. Rothkopf, and S. S. Smith, "Reliability and inventory in a production storage system," *Management Sci.*, vol. 25, pp. 799-807, 1979.
- [12] A. Federgruen and P. Zipkin, "An inventory model with limited production capacity and uncertain demands. I: The average cost criterion," *Math. Op. Res.*, vol. 11, pp. 193-207, 1986.
- [13] M. C. Fu and J. Q. Hu, "On the relationship of capacitated production/inventory models to manufacturing flow control models," *Op. Res. Lett.*, 1995, to appear.
- [14] J. Q. Hu and D. Xiang, "A queuing equivalence to optimal control of a manufacturing system with failures," *IEEE Trans. Automat. Contr.*, vol. 38, pp. 499-502, 1993.
- [15] J. M. Cohen, *The Single Server Queue*, Amsterdam: North Holland, 1982.
- [16] H. Chen and D. D. Yao, "A fluid model for systems with random disruptions," *Op. Res.*, vol. 40 (supp. 2), pp. S239-S247, 1992.
- [17] O. Kella and W. Whitt, "A storage model with a two state random environment," *Op. Res.*, vol. 40 (supp. 2), pp. S257-S262, 1992.
- [18] J. Buzacott and J. Shanthikumar, *Stochastic Models for Manufacturing Systems*, Englewood Cliffs, NJ: Prentice Hall, 1993.
- [19] L. Kleinrock, *Queueing Systems I*, New York: Wiley, 1975.

Book Reviews

In this section, the IEEE Control Systems Society publishes reviews of books in the control field and related areas. Readers are invited to send comments on these reviews for possible publication in the Technical Notes and Correspondence section of this TRANSACTIONS. The CSS does not necessarily endorse the opinions of the reviewers.

If you have used an interesting book for a course or as a personal reference, we encourage you to share your insights with our readers by writing a review for possible publication in the TRANSACTIONS. All material published in the TRANSACTIONS is reviewed prior to publication. Submit a completed review or a proposal for a review to:

D. S. Naidu
Associate Editor—Book Reviews
College of Engineering
Idaho State University
833 South Eighth Street
Pocatello, ID 83209

Analysis and Control of Nonlinear Infinite Dimensional Systems
— V. Barbu. (San Diego, CA: Academic Press, 1993). Reviewed by
H. O. Fattorini.

I. DIFFERENTIAL EQUATIONS IN LINEAR SPACES

Let Ω be a domain with boundary Γ in m -dimensional Euclidean space \mathbb{R}^m . An initial or initial-boundary value problem such as

$$u_t(t, x) = \Delta u(t, x), \quad u(0, x) = u_0(x) \quad (x \in \Omega) \quad (1)$$

$$u(x, t) = 0 \quad (x \in \Gamma) \quad (2)$$

($x = (x_1, x_2, \dots, x_m)$, Δ the Laplacian in the space variables) can be studied in the form of an "ordinary differential" evolution equation

$$u'(t) = Au(t), \quad u(0) = u_0 \quad (3)$$

in a suitable function space E ; $u(t)$ takes values in E and $A = \Delta$, the boundary conditions included in the definition of the domain $D(A)$ of A or of the space E . For instance, we may take $E = L^2(\Omega)$, $D(A)$ consisting of all $u \in L^2(\Omega)$ such that $\Delta u \in L^2(\Omega)$ (in the sense of distributions) and $u = 0$ on Γ . Alternately, we may set $E = C_0(\bar{\Omega})$ (the space of all continuous functions on the closure $\bar{\Omega}$ that vanish at the boundary Γ); in this case, the boundary condition (2) is included in the definition of the space rather than that of $D(A)$.

The formal similarity of (3) with a linear, constant coefficient system of ordinary differential equations gives insight into the original problem (1)–(2); for instance, we may expect to be able to write the solution in the form $u(t) = S(t) = \exp(tA)u_0$, which is true if the exponential is correctly interpreted. This approach to evolution problems for partial differential equations was initiated by Hille [9] and Yosida [20] under the semigroup theory point of view of studying the exponential operator equation $S(s+t) = S(s)S(t)$, $S(0) = I$; their main result, the Hille–Yosida theorem, characterizes the infinitesimal generator $A = \lim_{h \rightarrow 0} h^{-1}(S(h) - I)$ of $S(t)$ in terms of the resolvent operator $(\lambda I - A)^{-1}$. Equivalently, it characterizes the operators A (not necessarily bounded or everywhere defined) such that the initial value problem (3) is well posed; this means solutions exist for a dense set of initial data and depend

continuously on these initial data. (The Hille–Yosida theorem was given full generality by R. S. Phillips, I. Miyadera, and W. Feller). Among the many subsequent highlights of the linear theory, we mention the introduction of dissipative operators in Hilbert space [19] (corresponding to an important particular case of the Hille–Yosida Theorem) and their Banach space extension in [16].

Semigroup theory was extended to nonlinear operators by Komura in [11] in Hilbert spaces, soon after by Kato [10] in certain "smooth" Banach spaces, and finally by Crandall and Liggett [5] in arbitrary Banach spaces (see [17] for more information). A class of equations (2) lying between the linear and the fully nonlinear is that of semilinear and quasilinear equations; these can be treated by a combination of linear semigroup methods combined with results on integral equations and fixed point theory. See [8] and [18] for more information on these equations and on linear semigroup theory.

The author of the present book has contributed many important results on nonlinear semigroup theory as well as the first textbook in this area [1] (the Romanian version appeared in 1974). Part of this work can be considered as an update of [1]. It includes a short review of linear semigroup theory, the key results of nonlinear semigroup theory, and numerous applications to partial differential equations.

II. INFINITE DIMENSIONAL OPTIMAL CONTROL THEORY

Finite dimensional control theory deals with systems described by ordinary differential equations; infinite dimensional control theory studies systems described by partial differential equations (and also by functional differential equations). Control theory of partial differential equations began to be practiced by engineers in the fifties, and was raised to a higher mathematical level in the Soviet Union by Butkovskii and others during the sixties [3]. Using a more abstract approach, [15] brought to bear on control theory the extensive theory of boundary value problems developed in the previous decade by Magenes and the author. At about the same time, [6], [7] a different approach to evolution control problems was initiated, based on the Banach space version (3) of the equations. Since then, mathematical optimal control theory of partial differential equations has developed along these two lines, sometimes combining both approaches. There are problems (such as those described by elliptic equations or elliptic variational inequalities) where the evolution approach (3) is not germane; however, many other tools of functional analysis apply.

The reviewer is with the Department of Mathematics, University of California, Los Angeles, CA 90024-1555 USA.

IEEE Log Number 9408268.

For several years the emphasis was on obtaining versions of Pontryagin's maximum principle for optimal control problems of all types. Recently, much work has been done on the dynamic programming approach to control of partial differential equations which extends the finite dimensional theory in [2]. Potentially the most promising approach (it provides closed-loop feedback solutions), it was partly heuristic even in finite dimensions. New developments in nonlinear analysis such as viscosity solutions of nonsmooth nonlinear partial differential equations have plugged many of the mathematical gaps and put the theory on a firmer footing, although much remains to be done.

The present book comprises both control of steady state and evolution systems. Among the first are optimal control of elliptic equations and variational inequalities, including the Signorini problem, the obstacle problem and free boundary problems. Treatment of evolution problems includes both results of maximum principle type and the dynamic programming approach.

We mention in passing that the 'engineering' control theory of partial differential equations has gone its own way although often intersecting with the mathematical theory, it usually approaches problems by discretization and application of finite dimensional theory, sometimes leaving aside rigorous proofs. On this it must be said, however that mathematics lags behind the applications in this area as in others. There are works such as [4] that straddle both fields: they adopt in some ways the engineering point of view but provide mathematical rigor as well.

III. CONCLUSIONS

Infinite dimensional optimal control theory within acceptable mathematical standards requires a considerable overhead. Without counting measure theory, a great deal of partial differential equations and functional analysis must be absorbed before one can face even the simplest problems. The same can be said of the abstract theory of evolution equations. This book provides a clear example filled with exploration into the two subjects written by a main protagonist in the latest developments. The work can be used as both a textbook in infinite dimensional control theory or in a modern course in partial differential equations presenting much of the recent research into such subjects as the dynamic programming approach. We believe its main audience will consist of mathematicians but there are in many advanced graduate students in engineering that could profit from it. It is a welcome contribution in a field not very well represented. There are other works on the market mostly in such other subjects

as stabilization by feedback [4], [13], the linear-quadratic problem [14], or controllability [12]. This book is a welcome and timely addition to the literature.

REFERENCES

- [1] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Leyden-Noordhoff, 1976.
- [2] R. Bellman, *Dynamic Programming*, Princeton, NJ: Princeton Univ. Press, 1957.
- [3] A. G. Butkovskii, *Optimal Control of Distributed Parameter Systems*, Moscow: Nauka, 1965.
- [4] G. Chen and J. Zhou, *Vibration and Damping in Distributed Systems*, vol. I, II, Boca Raton, FL: CRC Press, 1993.
- [5] M. Ciandali and T. Liggett, "Generation of semi groups of nonlinear transformations on general Banach spaces," *Amer. J. Math.*, vol. 93, pp. 263-278, 1971.
- [6] Y. I. Egorov, "Necessary conditions for the optimal control in Banach spaces," *Math. Sbornik*, vol. 64, pp. 79-101, 1964.
- [7] H. O. Fattorini, "Time optimal control of solutions of operational differential equations," *SIAM J. Contr.*, pp. 54-59, 1964.
- [8] —, *The Cauchy Problem*, Cambridge: Cambridge Univ. Press, 1983.
- [9] F. Hille, *Functional Analysis and Semi Groups*, New York: Amer. Math. Soc., 1948.
- [10] T. Kato, "Nonlinear semigroups and evolution equations," *J. Math. Soc. Japan*, vol. 19, pp. 508-520, 1967.
- [11] Y. Komura, "Nonlinear semigroups in Hilbert space," *J. Math. Soc. Japan*, vol. 19, pp. 493-507, 1967.
- [12] W. Krabs, *On Moment Theory and Controllability of One Dimensional Vibrating Systems and Heating Processes* (Springer Lecture Notes in Control and Information Sciences), vol. 173, Berlin: Springer-Verlag, 1992.
- [13] J. Lagnese, "Boundary stabilization of thin plates," *SIAM Studies in Appl. Math.*, 1989.
- [14] I. Lasiecka and R. Triggiani, *Differential and Algebraic Riccati Equations with Application to Boundary Point Control Problems, Continuous Theory and Approximation Theory* (Springer Lecture Notes in Control and Information Sciences), vol. 164, Berlin: Springer-Verlag, 1991.
- [15] J. L. Lions, *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles*, Paris: Dunod, 1968.
- [16] G. Lumer and R. S. Phillips, "Dissipative operators in a Banach space," *Pac. J. Math.*, vol. 11, pp. 670-698, 1961.
- [17] I. Miyadera, *Nonlinear Semigroups*, New York: Amer. Math. Soc., 1992.
- [18] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, New York: Springer, 1983.
- [19] R. S. Phillips, "Dissipative operators and parabolic partial differential equations," *Commun. Pure Appl. Math.*, 12, pp. 249-276, 1959.
- [20] K. Yosida, "On the differentiability and the representation of one parameter semigroups of linear operators," *J. Math. Soc. Japan*, vol. 1, pp. 15-21, 1948.

